Precise Object Localization using eXplainable AI Methods

Institutions

1 UNIVERSITÀ DI TORINO https://www.unito.it/

2 RuleX https://www.rulex.ai/

PhD Student, Cycle 38
Muhammad Rashid[1]

Academic Supervisor:
Elvio G. Amparore[1]

Industrial Supervisors:
Enrico Ferrari[2]
Damiano Verda[2]

Fundings

NEXT PERCEPTION https://nextperception.eu/

DistriMuSe

HORIZON-KDT-JU-2023-2-RIA | 2023-2026

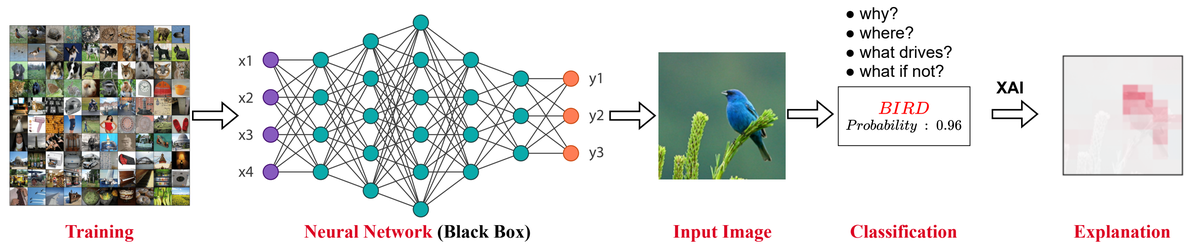
eXplainable AI (XAI)

Introduction

- Why and What is eXplainable Artificial Intelligence(**XAI**)
- **Semantic Analysis** of Machine Learning Models
- Explanations from Image Classification Models
- Explanations from **Anomaly Detection** Systems
- **Class Activation Map** based methods
- **Better evaluation** of AI systems using **XAI**
- **Precise Object Localization** in Multiple Application

Research Questions & Solutions

- Under representation of synthetic data (RQ1)
- Anomaly map finds precise Localization of Anomaly (RQ2)
- Shapely Values generation care about data? (RQ3)
- Anomaly Score is best choice for decision? (RQ4)
- SoTA evaluation metrices perform faithful evaluation? (RQ5)

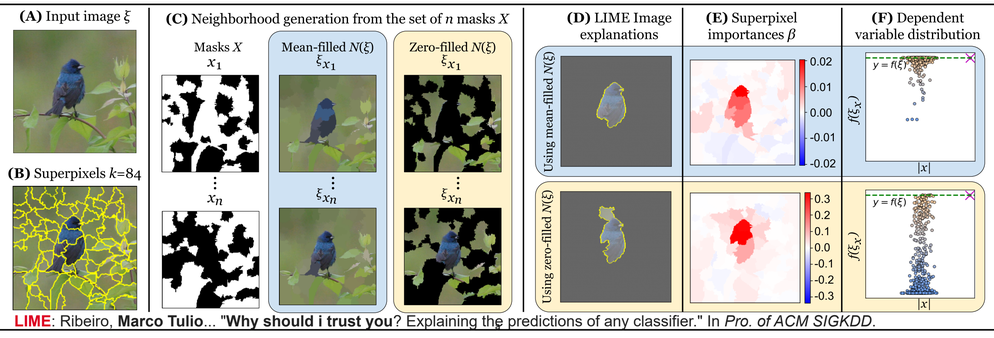


RQ1: Under Representation of Synthetic Neighborhood

(a) Under Representation of Synthetic Neighborhood

- Perturbation based methods use synthetic neighborhood.
- Does SoTA XAI methods **represent** 100% of synthetic neighborhood?
- Under-representation causes confused and meaningless explanations.

(b) SoTA



LIME: Ribeiro, **Marco Tulio**... "**Why should i trust you**? Explaining the predictions of any classifier." In *Pro. of ACM SIGKDD*.

(c) Proposed

Masks Generation

mask vector x are sampled

$x[i] \sim B(0.5), \quad 1 \le i \le k$

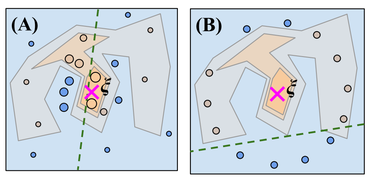
Dependent Variables

$Y = \{f(\xi_x) \mid \xi_x \in N(\xi)\}$

Distance Function

ξ should weight more

$w_x = \exp\left(\frac{-D(x)^2}{\sigma^2}\right)$

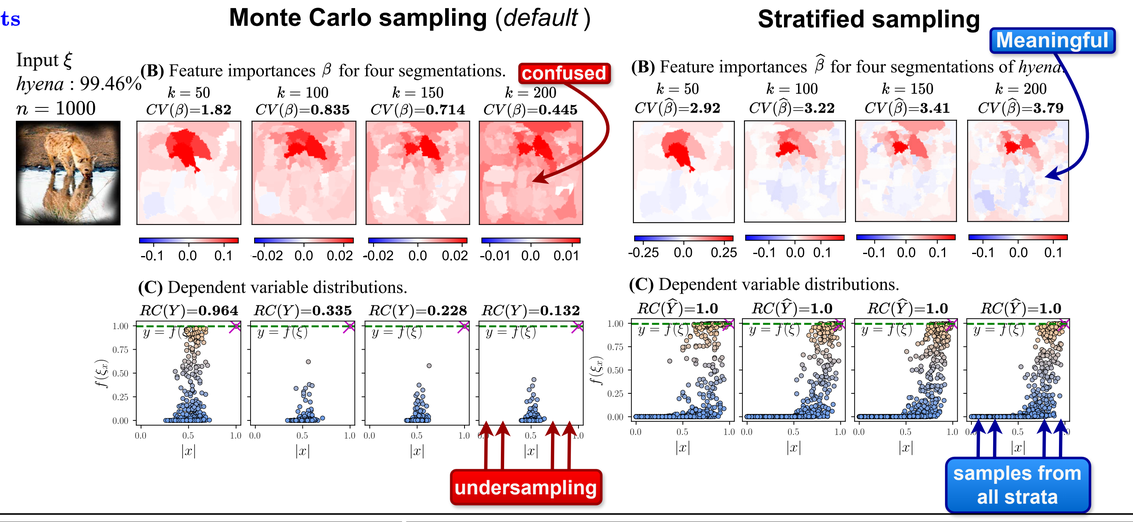


Stratified sampling

- Partitioning stratified on the number of masked superpixels.
- $\mathcal{X}^{(i)}$ = set of possible masks having $|x| = i$
- Stratum i size is known a priori:

$|\mathcal{X}^{(i)}| = \binom{k}{i}, \quad 0 \le i \le k$

(d) Results



(e) Experimental Evaluation

- $CV(\beta) = \frac{\sigma_\beta}{\mu_\beta}$ • $RC(Y) = \frac{IQR_{1-99}(Y)}{f(\xi)}$

Conclusions

- Reformulation of LIME sampling strategy.
- Practical improvements, at no additional cost.

The 38th Annual **AAAI** Conference on Artificial Intelligence

Using Stratified Sampling to Improve LIME Image Explanations

Muhammad Rashid[1], Elvio G. Amparore[1], Enrico Ferrari[2], Damiano Verda[2]

Source Code

- pip install lime_stratified
- github/rashidrao-pk/lime_stratified

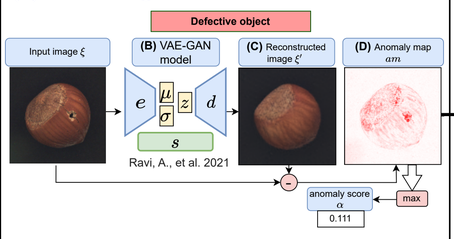
RQ2: Anomaly map finds precise Localization of Anomaly? Integration of XAI for Anomalies in Images

(a) Problem Statement

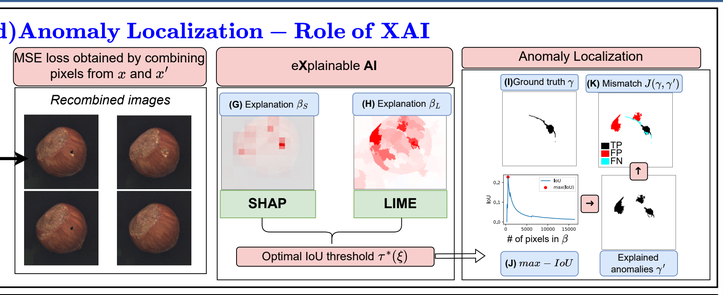
Separating real anomaly and background noise (if any)

- Case study Explaining Anomaly Detection(***AD***) Systems
- Verifying anomaly detection systems
- Precise anomaly localization using XAI
- Use of One Class Classification based Self Supervised Learning for Anomaly Detection

(c) Detecting Anomalies



(d) Anomaly Localization – Role of XAI



(f) Conclusions

- XAI methods are relevant in finding the **true drivers** behind AI systems using techniques like classification and/or anomaly detection.
- Case study based on **reconstruction error maps** generated from **VAE-GAN** models.
- **Multiple XAI** techniques to separate the **reconstruction error** (noise) from the **anomaly** (if any).
- A sample may be detected as **anomalous** for the **wrong reasons**, yet this misbehaviour may not be detectable from the information provided by the anomaly detection system alone → Role of XAI!

(b) Training of Variational AutoEncoder Generative AI (VAE – GAN)

Encoding

- $z = e(\xi)$

Decoding

- $\xi' = d(z)$

Anomaly Map

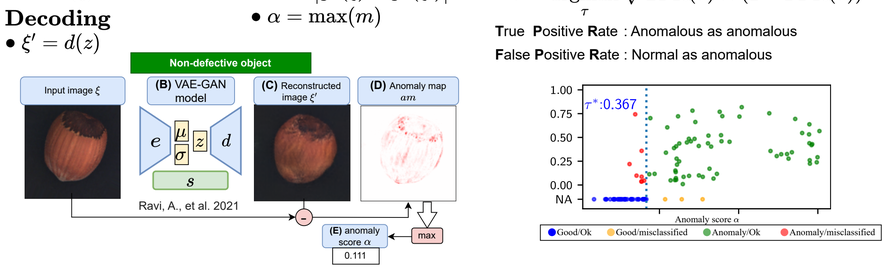
- $m = |gs(\xi) - gs(\xi')|$
- $\alpha = \max(m)$

Global Threshold

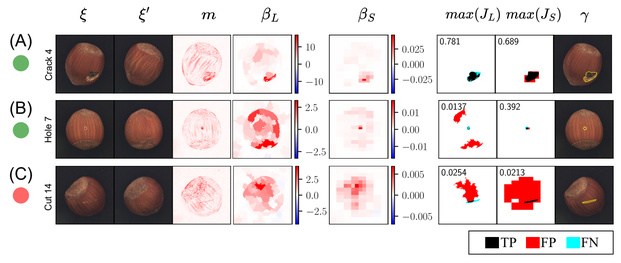
- $\tau = \underset{\tau}{\text{argmax}} \sqrt{\text{TPR}(\tau) \times (1 - \text{FPR}(\tau))}$

True **P**ositive **R**ate : Anomalous as anomalous

False **P**ositive **R**ate : Normal as anomalous



(e) Results



*SHAP -Scott, M.. **A unified approach to interpreting model predictions**.

*LIME: **Marco Tulio**. "**Why should i trust you**? Explaining the predictions of any classifier."

2nd World Conference on e**X**plainable **A**rtificial **I**ntelligence

Can I Trust my Anomaly Detection system? A case study based on eXplainable AI

Muhammad Rashid[1], Elvio G. Amparore[1], Enrico Ferrari[2], Damiano Verda[2]

Source Code

github/rashidrao-pk/anomaly_detection_trust_case_study

RQ3: Shapely Values generation care about data?

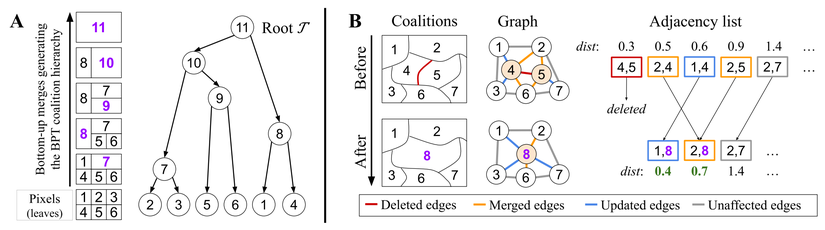
(a) Precise Object Localization

- Does SoTA *XAI* methods provide precise object localization?
- ***Binary Hierarchy*** of ***Owen Approximation*** of ***Shapley Values***
- Use of **Data Aware approach -** Binary Partition Tree (***BPT***)
- **Feedback loop** between **Segmentation** and **Shapely Values Generation**

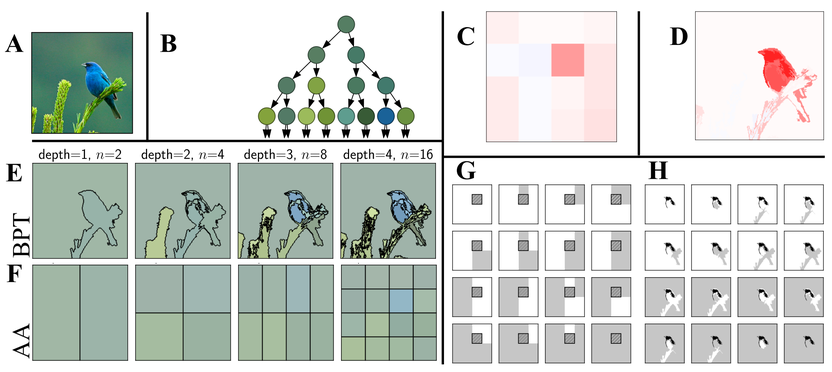
$$\Delta_i(S) = \nu(S \cup \{i\}) - \nu(S)$$

(b) How Binary Partition Tree(BPT) works on Image?

$$dist(T_i, T_j) = dist^2_{\text{color}}(T_i, T_j) \cdot area(T_i, T_j) \cdot \sqrt{perim(T_i, T_j)}$$

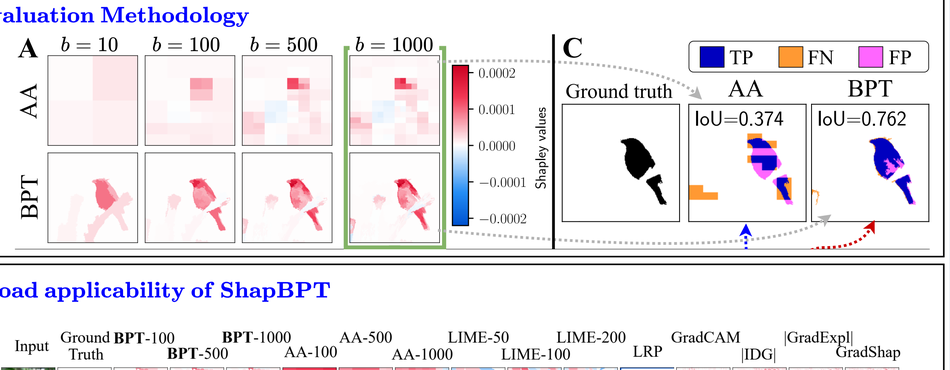


(c) Comparison of BPT & Axis Aligned(AA) Hierarchies

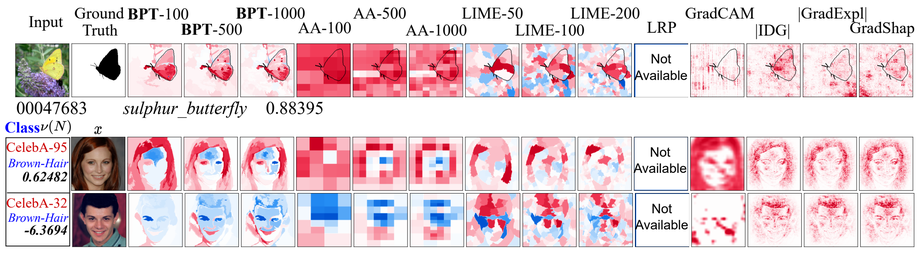


*AA: SHAP -Scott, M.. **A unified approach to interpreting model predictions**.

(d) Evaluation Methodology



(e) Broad applicability of ShapBPT



(f) Performance Measures

$$AUC^+ = \int_0^1 \nu(S^{[q]})\,dq$$
$$AUC^- = \int_0^1 \nu(\mathcal{N} \setminus S^{[q]})\,dq$$

$$MSE^+ = \int_0^1 \left(\nu(S^{[q]}) - \eta(S^{[q]})\right)^2 dq,$$
$$MSE^- = \int_0^1 \left(\nu(\mathcal{N} \setminus S^{[q]}) - \eta(\mathcal{N} \setminus S^{[q]})\right)^2 dq$$

$$AUIOU = \int_0^1 J(S^{[q]}, G)dq)$$
$$maxIoU = \max_{q \in [0,1]} \left(J(S^{[q]}, G)\right)$$

(g) Experimental Evaluation

- Performance Curve based evaluation
- GroundTruth based evaluation
- Controlled setup for exact IoU
- Multiple replacement values
- *ImagNet Dataset*
- *ResNet50, VGG16, Swin-ViT*
- *Multiple Applications*
  - *Object Localization* (***ImagNet***)
  - *Facial Attributes Localization* (***CelebA***)
  - *Anomaly Localization* (***MVTec***)

International **C**onference on **L**earning **R**epresentations- **ICLR-25**

Shapley Image Explanations With Data-Aware Binary Partition Trees

Muhammad Rashid[1], Elvio G. Amparore[1], Enrico Ferrari[2], Damiano Verda[2]

Research Trainings & Further Plans

Training Activities

- PhD Course Work (Completed)
- Conferences (**ECML**, **AAAI**-24, **XAI** World)
- External Courses (Oxford ML School, Google)
- Teaching Collaboration (Cyber-Security Course)
- *Served as **Reviewer** in A & B ranked conferences.*

Enterprise Time

First Half:

- Basic Training Courses
- Review, Learn and Implement
  - **One Class Classification**
  - **AD using Hand Crafted-Features**
  - **AD using Self Supervised Learning**

Second Half:

- **Example/Prototype** based explanations.
- Improvement of Anomaly Detection systems.
- Improved Anomaly Score Functions (**RQ4**).

Mobility Abroad

- Dept. of Robotics, University of Granada
- Generation of **Synthetic Data with Simulations**
- Conversion to **Realistic Data**
- Perform **Style Transfer**
- Use of ***Diffusion*** *Models*
- *Train **Anomaly Detection** System for **Robot**.*

ugr Universidad de Granada

Further Plans

- Better Performance Measures (**RQ5**)
- Practical Implications of XAI for Improved
  - Data Acquisition
  - Learning Patterns

Muhammad Rashid[1]
muhammad.rashid@unito.it
muhammad.rashid@rulex.it

Elvio G. Amparore[1]
elviogilberto.amparore@unito.it

Enrico Ferrari[2]
enrico.ferrari@rulex.ai

Damiano Verda[2]
damiano.verda@rulex.ai

Contact me

- LinkedIn/rashid-rao-cuipakistan