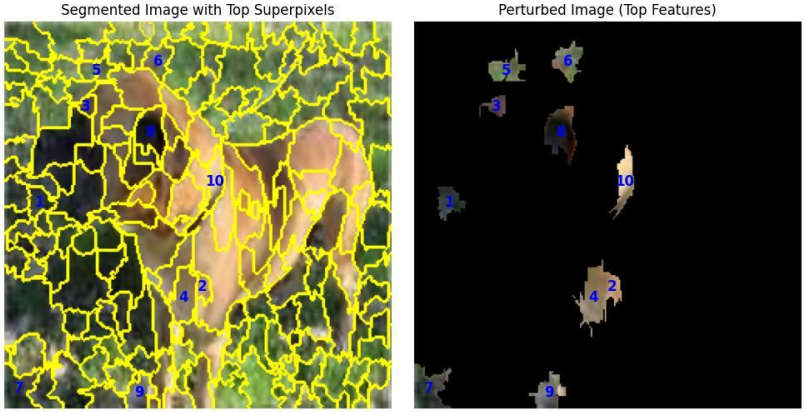
LIME

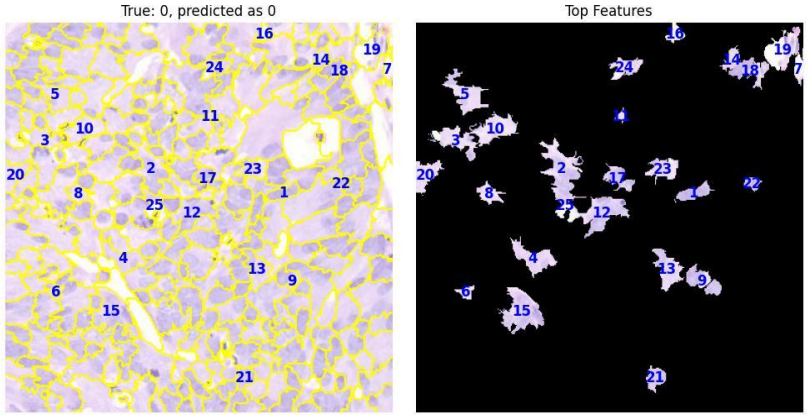
Rationale, Strengths, and Weaknesses

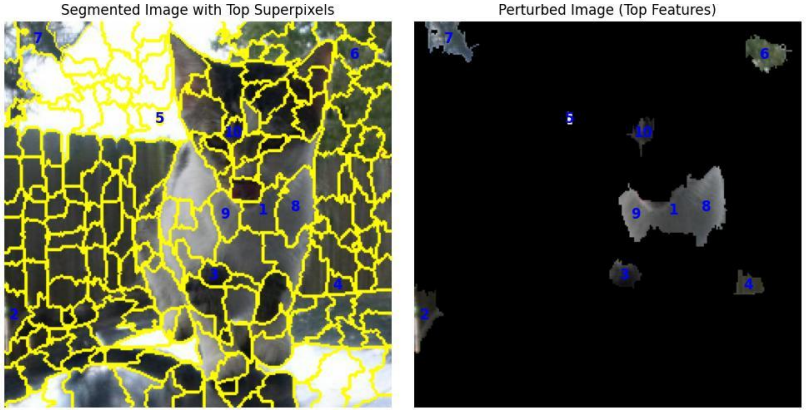
What is LIME?

LIME stands for **Local Interpretable Model-agnostic Explanations**.

It uses a surrogate model, surprisingly a **linear model**!



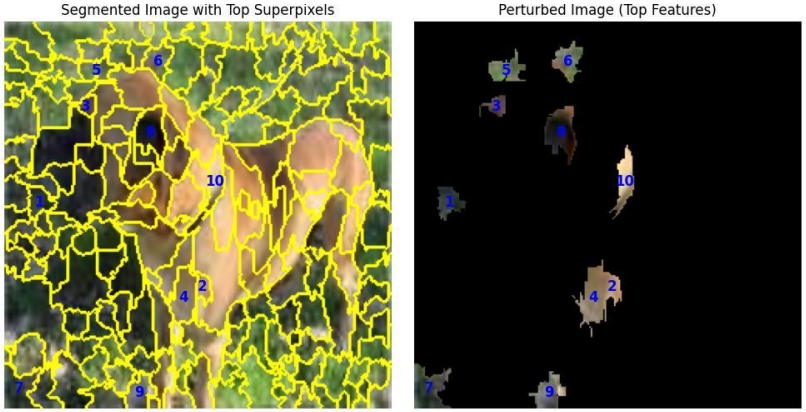


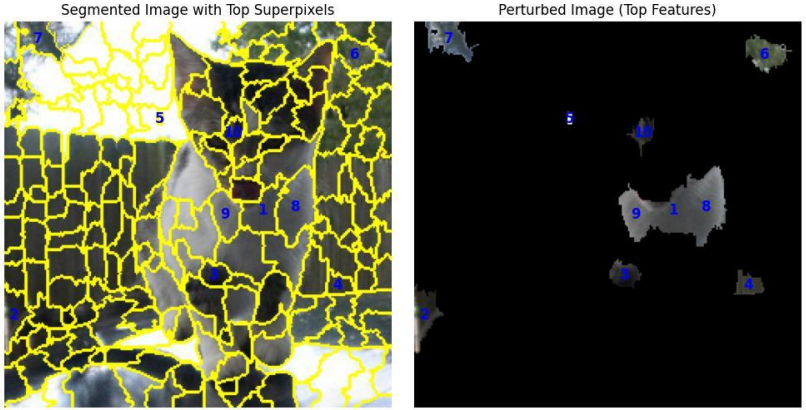


Rationale

Modern models like deep neural networks are **non-linear and complex**, often described as "black boxes" due to their lack of interpretability. Millions of para. LIME addresses this by:

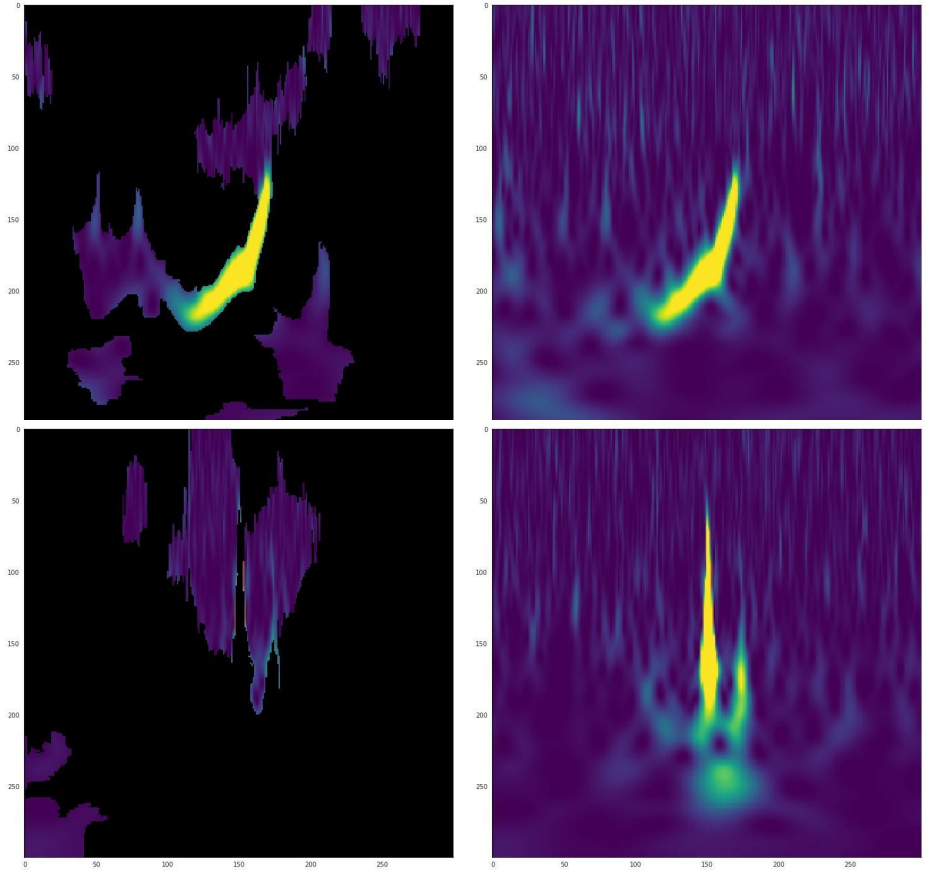
- Explaining individual predictions,
- Generating **local surrogate models** (usually a linear one) that approximate the black-box model **locally** (near a specific input),
- Enabling humans to **understand which features contributed most to a decision**.





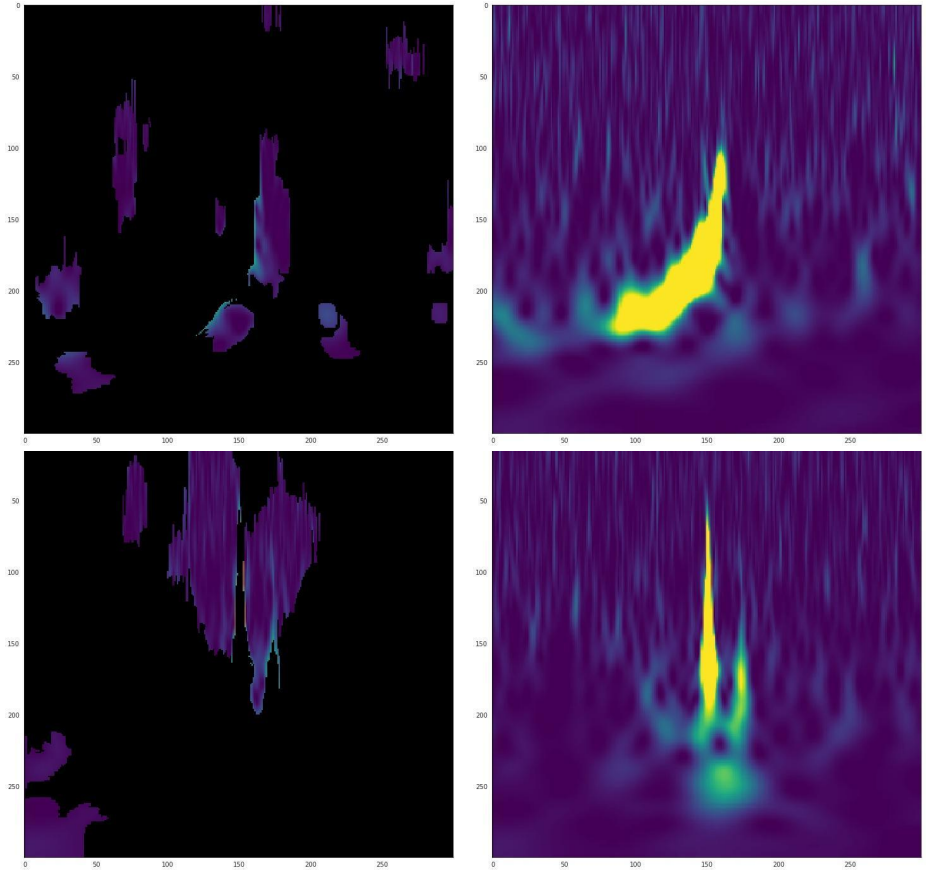
Strengths of LIME

- **Model-agnostic**: Works with **any classifier or regressor**, including deep neural networks.
- **Local fidelity**: Focuses on explaining **individual predictions**.
- **Human-understandable**: Produces intuitive visualizations for images and coefficients for tabular data.

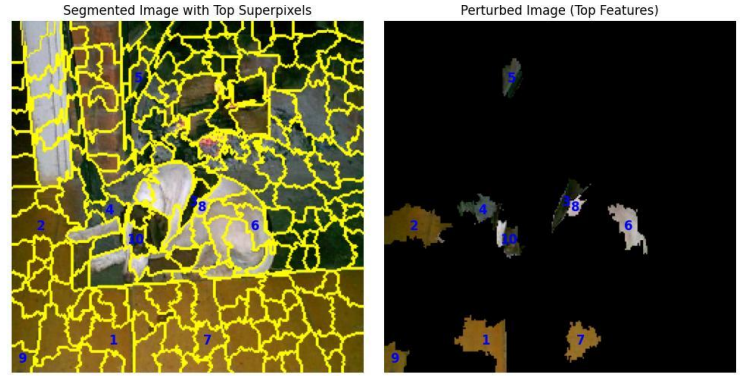


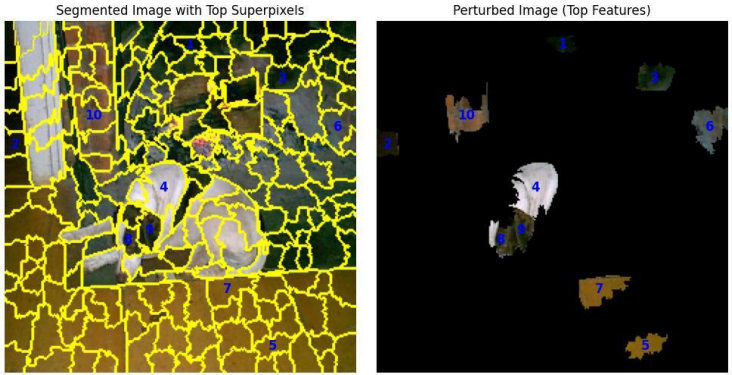
Strengths of LIME

- **Model-agnostic**: Works with **any classifier or regressor**, including deep neural networks.
- **Local fidelity**: Focuses on explaining **individual predictions**.
- **Human-understandable**: Produces intuitive visualizations for images and coefficients for tabular data.



Weaknesses of LIME

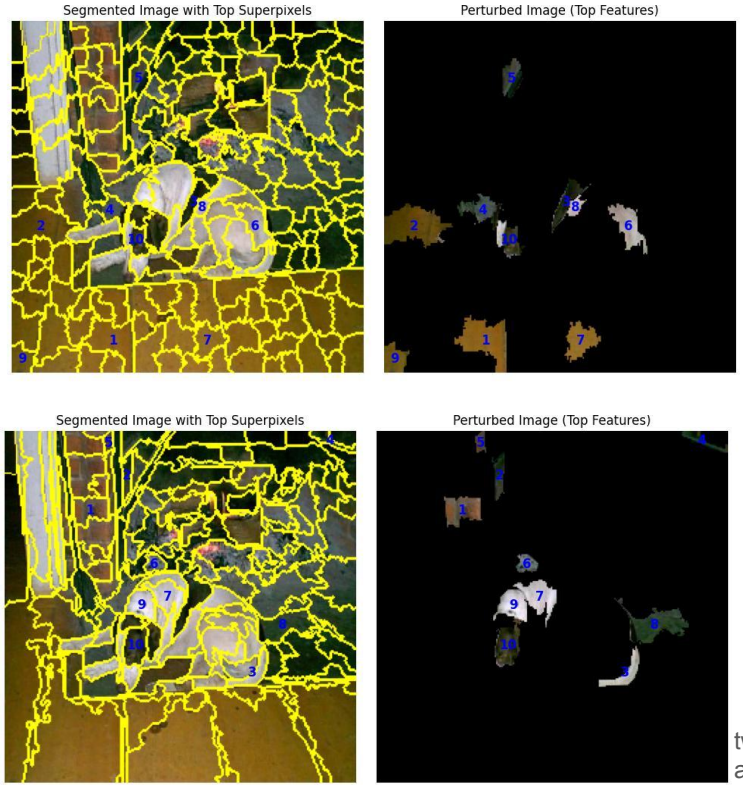




two independent runs, including training

- **Instability**: Different runs can produce different explanations due to randomness in perturbations.
- **Approximation Error**: The linear surrogate may not capture complex non-linear interactions, even locally.
- **Computational Cost**: Requires many queries to the black-box model for perturbations.
- **Superpixel Sensitivity**: Explanations depend on how the image was segmented; different segmentations can lead to different results.

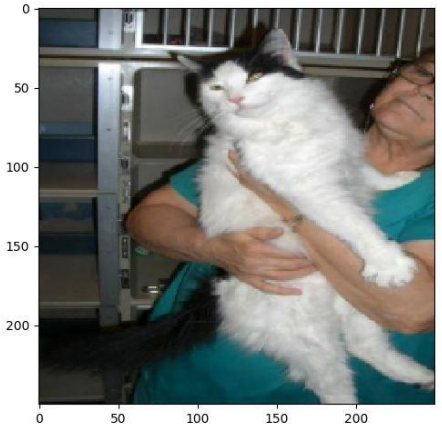
Weaknesses of LIME



two segmentation algorithms

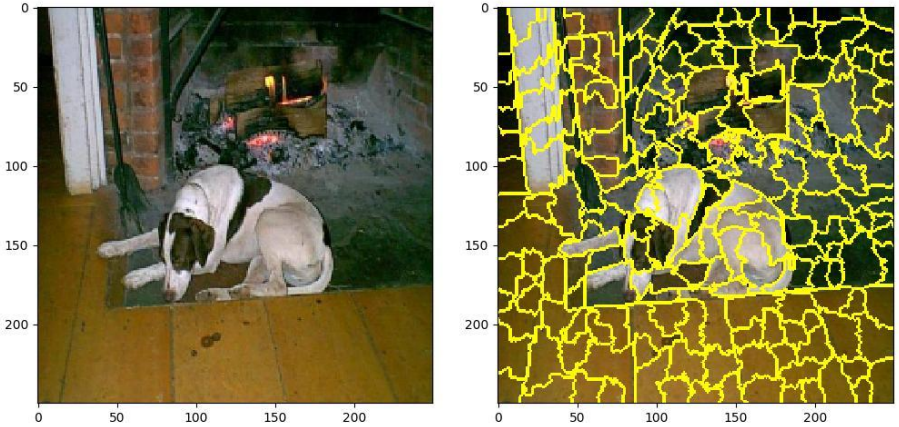
- **Instability**: Different runs can produce different explanations due to randomness in perturbations.
- **Approximation Error**: The linear surrogate may not capture complex non-linear interactions, even locally.
- **Computational Cost**: Requires many queries to the black-box model for perturbations.
- **Superpixel Sensitivity**: Explanations depend on how the image was segmented; different segmentations can lead to different results.

How LIME Works (Step-by-Step)

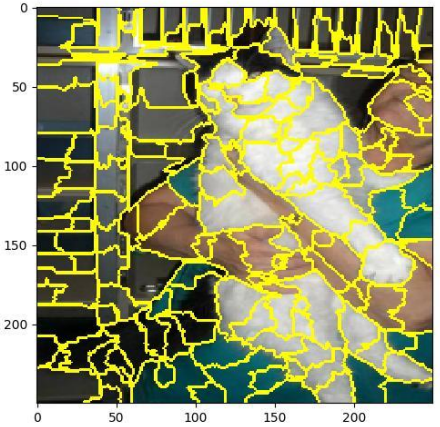


1. Input Image Segmentation

- The image is segmented into **superpixels** (groups of adjacent pixels with similar color or texture).



Each segment or superpixel is an “interpretable feature”.

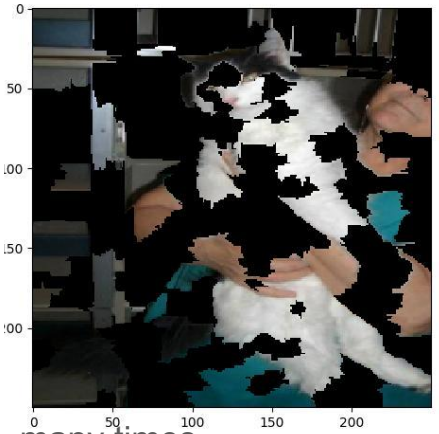


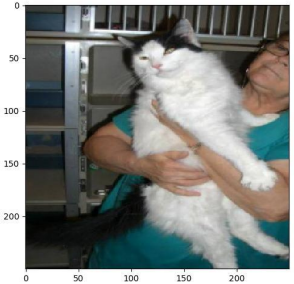
How LIME Works (Step-by-Step)

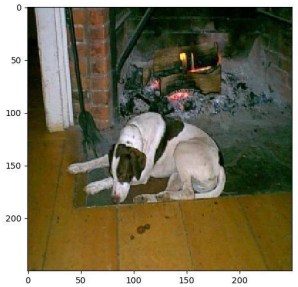
1. Input Image Segmentation

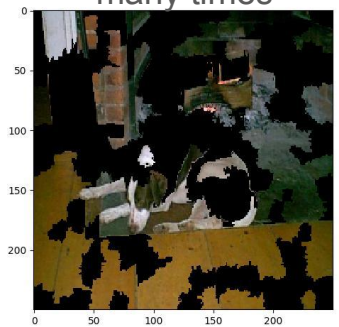
2. Perturbation

- Multiple perturbed versions of the original image are created by **randomly turning off some superpixels** (e.g., replacing them with black).
- Each perturbed image is fed into the **original model** to obtain predictions.
- These perturbations are just simple **binary representations** of the image indicating the “presence” or “absence” of each superpixel region.









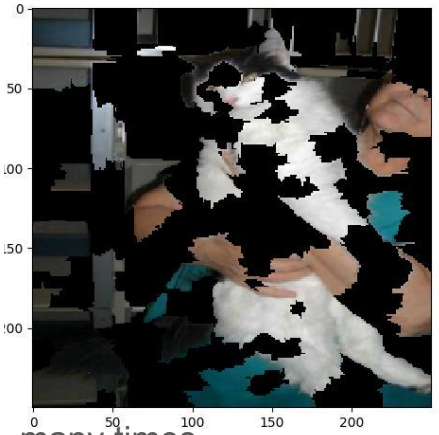
How LIME Works (Step-by-Step)

1. Input Image Segmentation

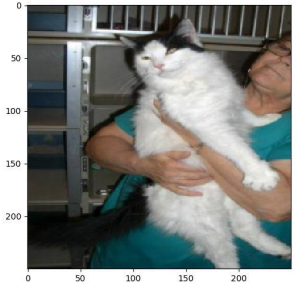
2. Perturbation

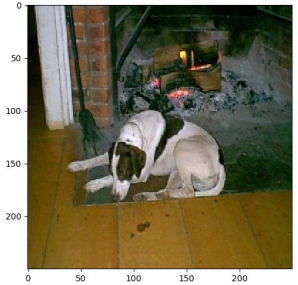
3. Weighting

- Perturbed images that are more similar to the original are assigned **higher weights**.
- This ensures that the **local approximation** focuses on the region around the original input (its neighbourhood).

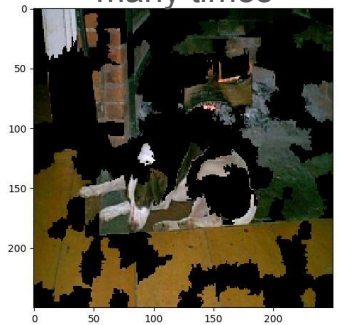


many times





many times



How LIME Works (Step-by-Step)

4. Fitting a Surrogate Model

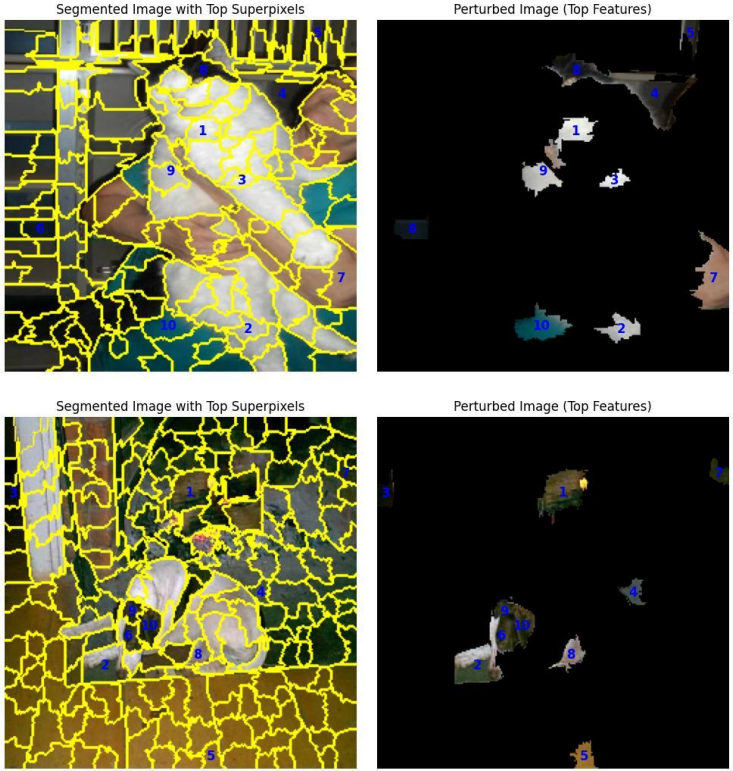
- A **simple interpretable model** (usually linear regression) is fitted on the data: the binary vector (which superpixels are present) vs. the prediction output of the black-box model.

  The (uninterpretable) complex model is approximated by a local (interpretable) more simple model.

- The **coefficients** of this surrogate model represent the **importance of each superpixel** for the prediction.

5. Generating the Explanation

- The top contributing superpixels are visualized as the **explanation**, showing which regions were most important for the model's prediction.

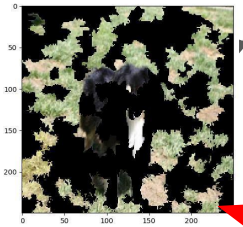


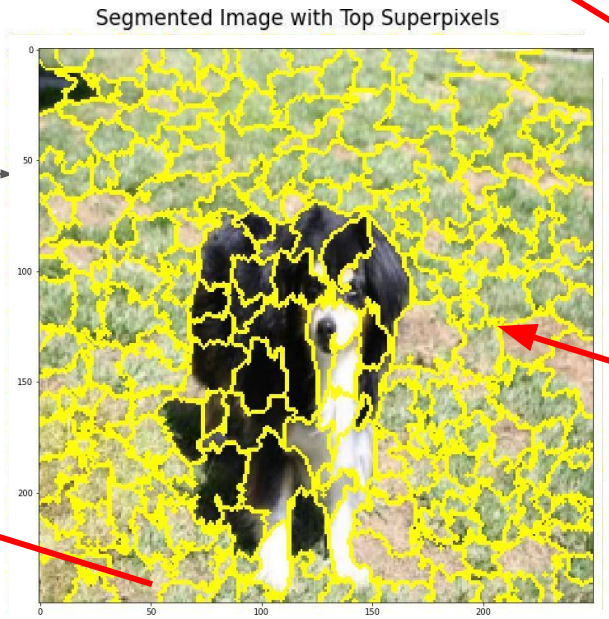
XAI, intuition and biased learning

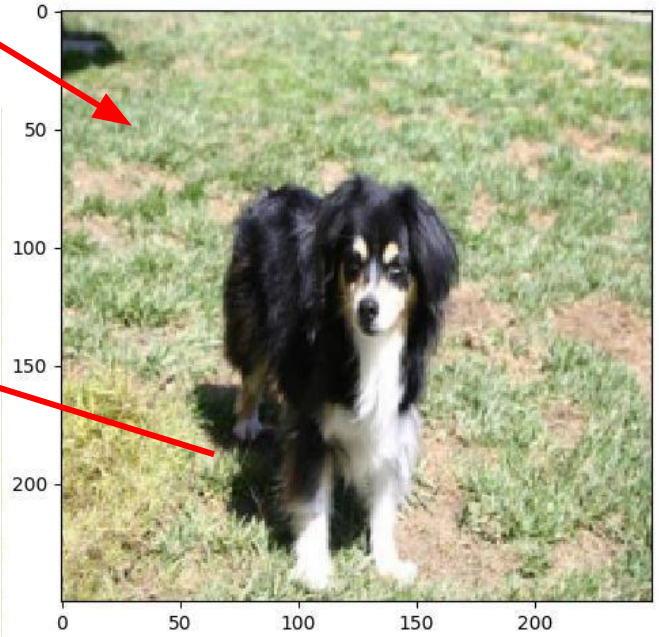
Dog correctly classified as dog (class 1): [4.8e-11 (cat), 1.0e+00 (dog)]

Image segmented in homogeneous patches (~250).

Randomly patches are switched on/off and then predicted.







Degradation of prediction illuminates the importance of the patches. And repeat …

Linear regressor with patches on/off as input and the performance degradation as output.

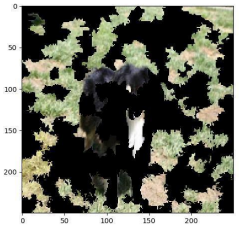
XAI, intuition and biased learning

Dog correctly classified as dog (class 1): [4.8e-11 (cat), 1.0e+00 (dog)]

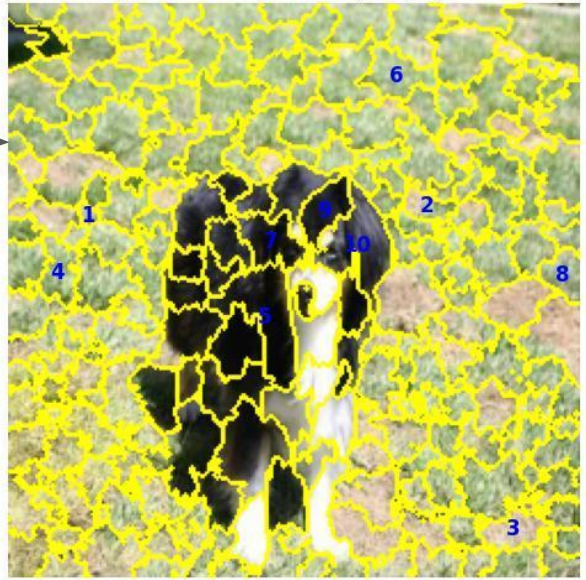
The 10 most important patches are mostly grass.

Image segmented in homogeneous patches (~250).

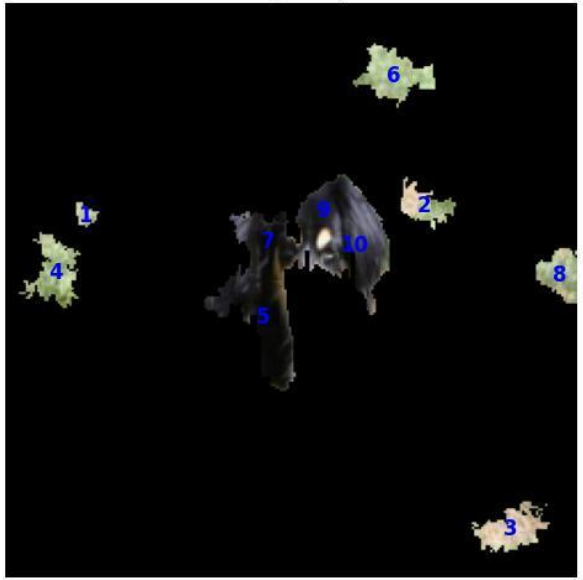
Randomly patches are switched on/off and then predicted.



Segmented Image with Top Superpixels



Perturbed Image (Top Features)



Degradation of prediction illuminates the importance of the patches. And repeat …

Linear regressor with patches on/off as input and the performance degradation as output.

Summary

LIME: Local Interpretable Model-agnostic Explanations.

Core

To explain an example, a surrogate model -–specifically a linear one— is created in the vicinity of the example by perturbing it.

The surrogate model approximates the complex dependency between input and prediction.

Advantages

Model-agnostic, local fidelity, easily human-understandable.

Limitations

Instability, approximation error, high computational cost, superpixel sensitivity.