Explainability

Eric Wong

9/29/2022

Local Linearity
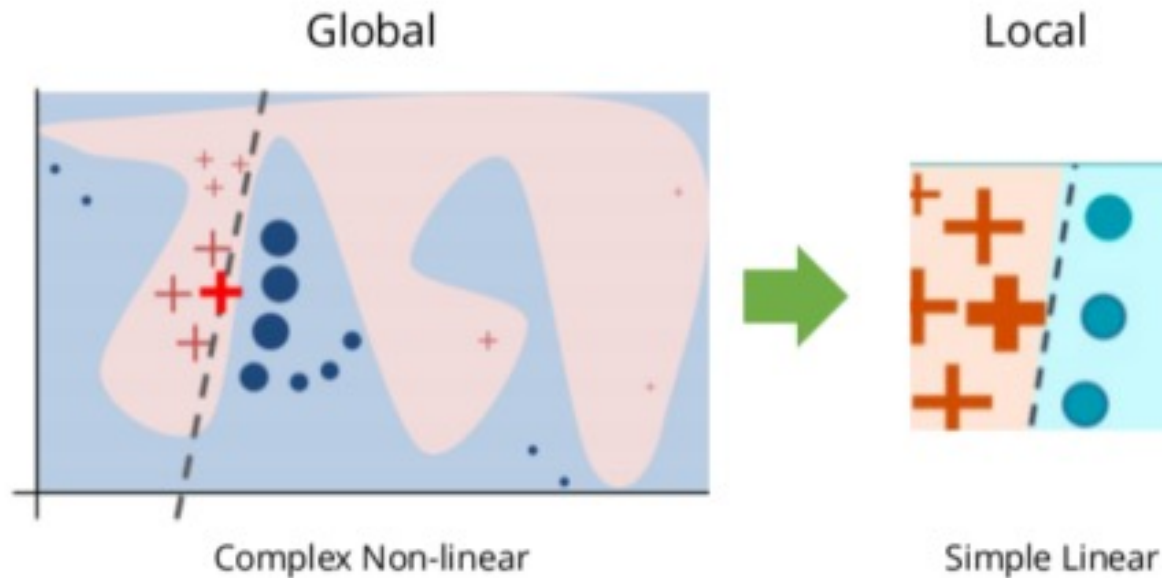
Global

Complex Non-linear

Local

Simple Linear

Marco Tulio Ribeiro "Local Interpretable Model-Agnostic Explanations (LIME): An Introduction"
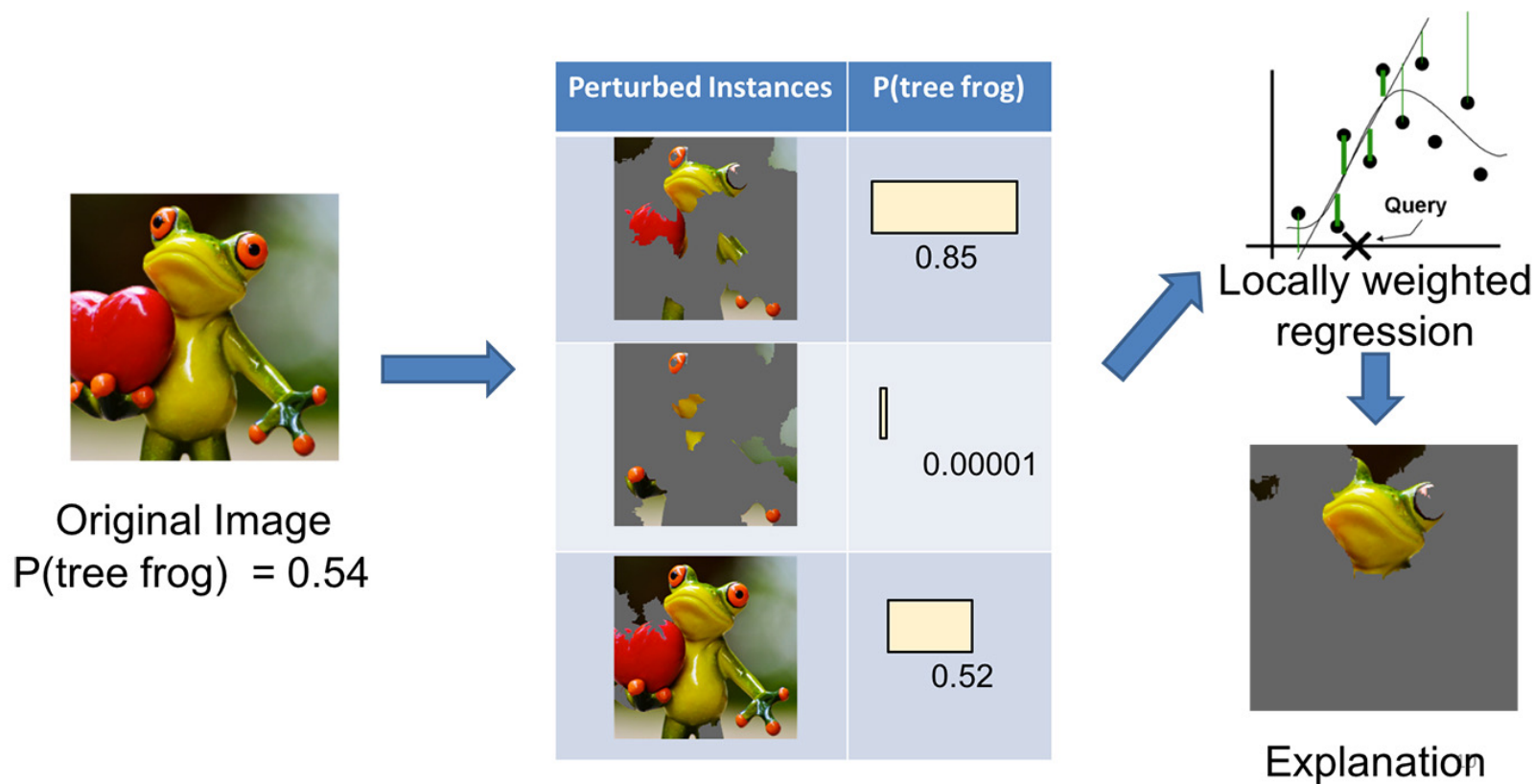
Superpixels for "interpretable" features

Original Image

Interpretable Components

Perturb superpixels

Original Image
P(tree frog) = 0.54

Perturbed Instances	P(tree frog)
	0.85
	0.00001
	0.52

Locally weighted regression

Query

Explanation

Marco Tulio Ribeiro "Local Interpretable Model-Agnostic Explanations (LIME): An Introduction"

Explaining images

Marco Tulio Ribeiro "Local Interpretable Model-Agnostic Explanations (LIME): An Introduction"

Explaining words

Marco Tulio Ribeiro "Local Interpretable Model-Agnostic Explanations (LIME): An Introduction"
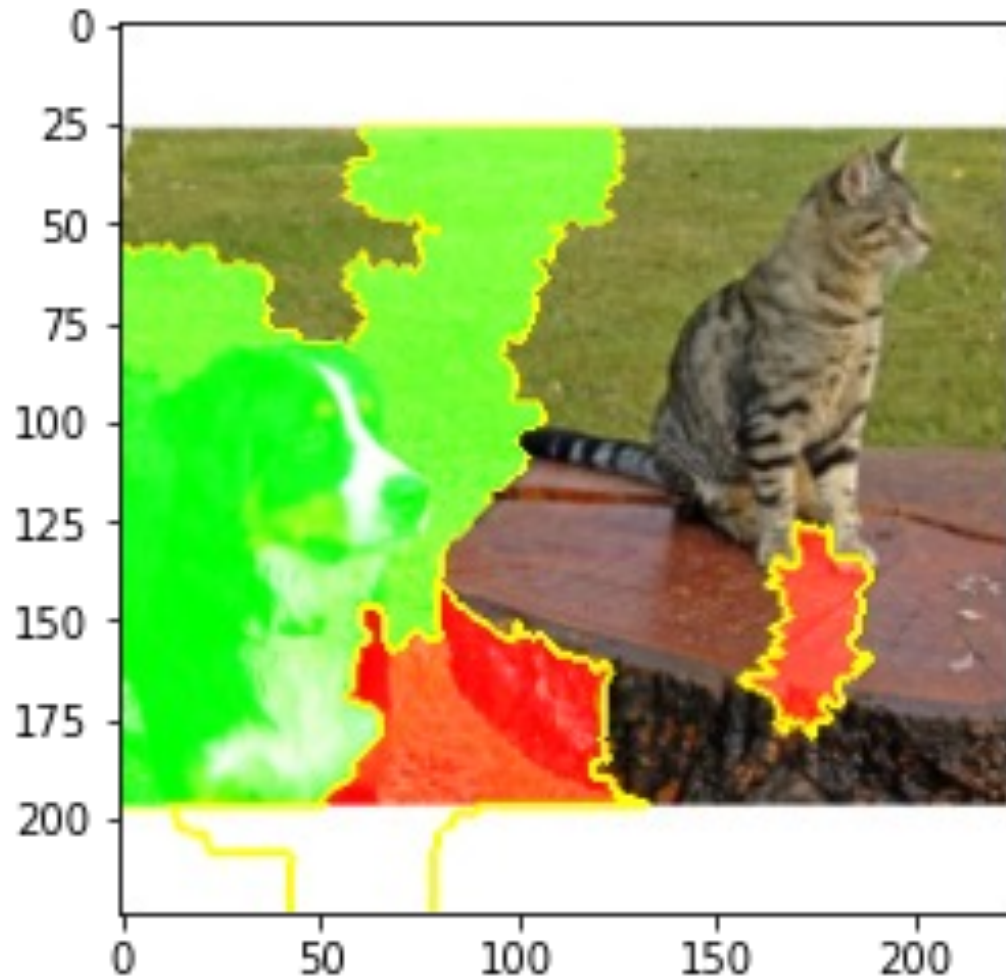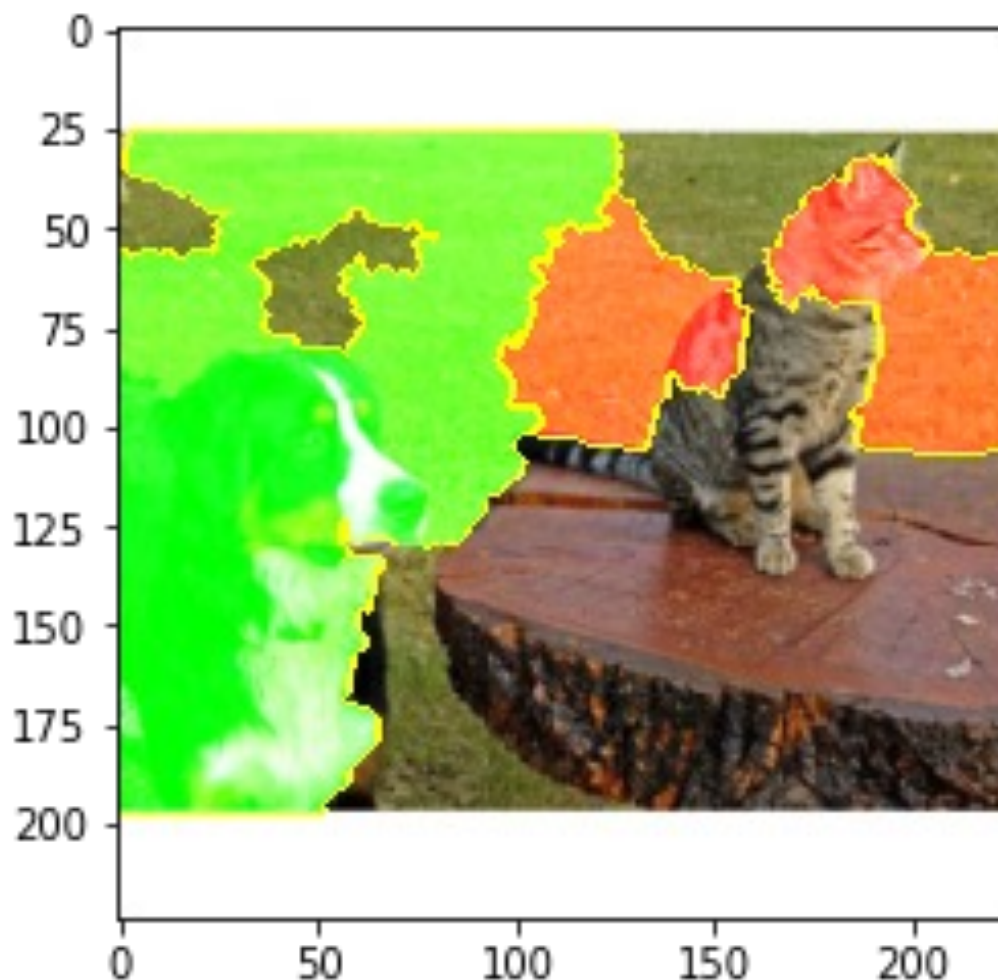
A closer look at dog
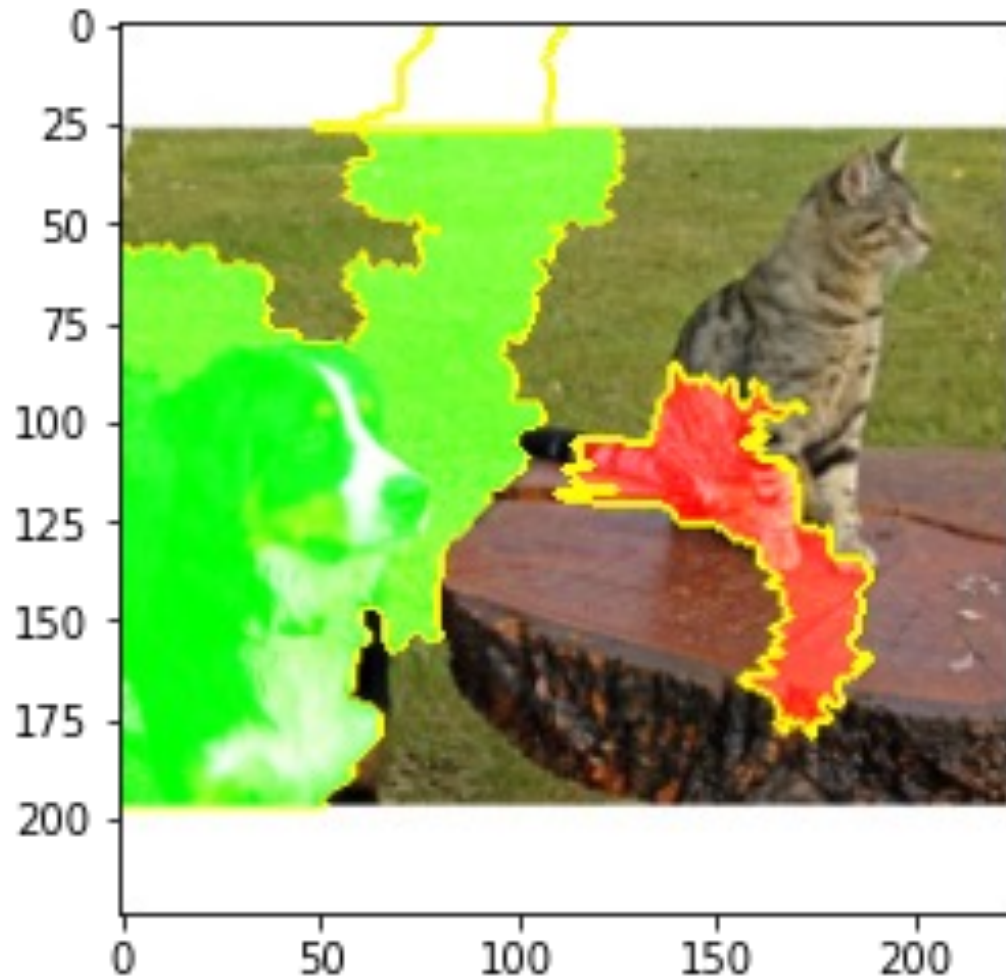
A closer look at dog

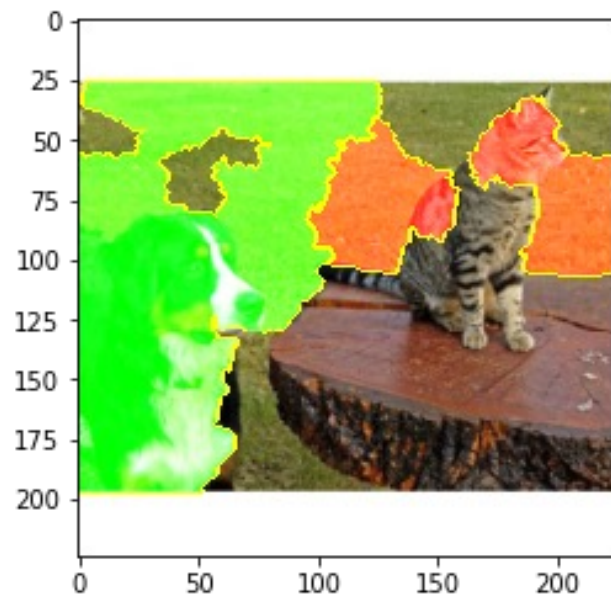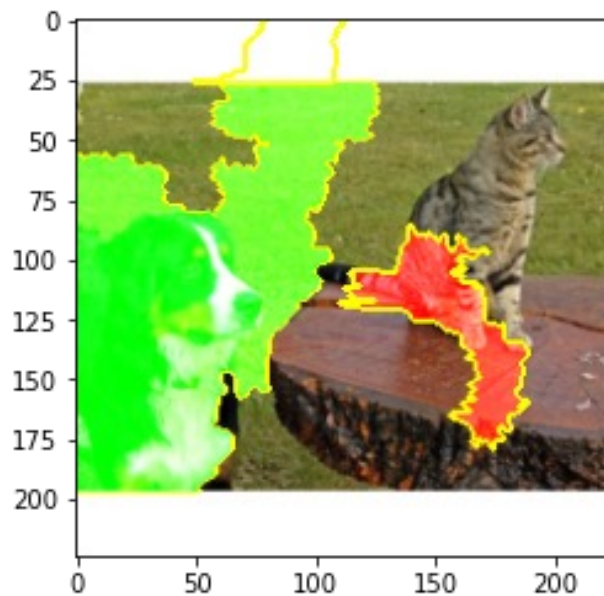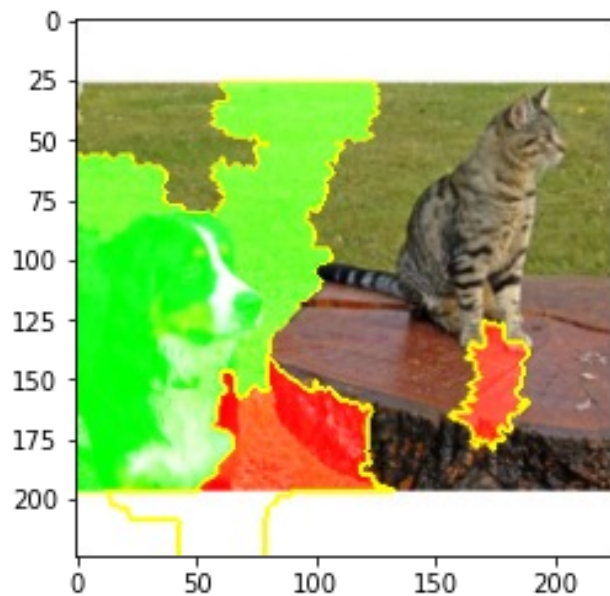# A closer look at dog

A closer look at dog

Local linearity?

Mask=0   Mask=0.5   Mask=1

Local linearity?