Do Explanations Explain?

Jacek Rutkowski

Wydział Matematyki, Informatyki i Mechaniki

January 2023

Problem formulation

- Which features are important? (feature attribution)
- –> But solutions differ!
- Which are correct?

Naive approach

- Compare the explanations against ground truth
- Problems with removing features

Axiomatic approach

- Explanations should satisfy some desirable properties
- We want to deal with particular implementations

Axioms

- Null-player axiom: if a player is null, he should have value zero
- Class-sensitivity: different outputs –> different explanations
- Feature-saturation

Goal of the paper

- Test explanations against axioms
- Construct environment to easily obtain ground truth attributions

Methodology: create ground truth

- Model is frozen
- We optimize the input
- We control how features contribute to the output

Null feature case

- Create two features (patches), f_a, f_{null}
- Two cases:
  - We take f_a: output should be a with and without f_{null}
  - We do not take f_a: output should be the same with and without f_{null}

Results for some methods

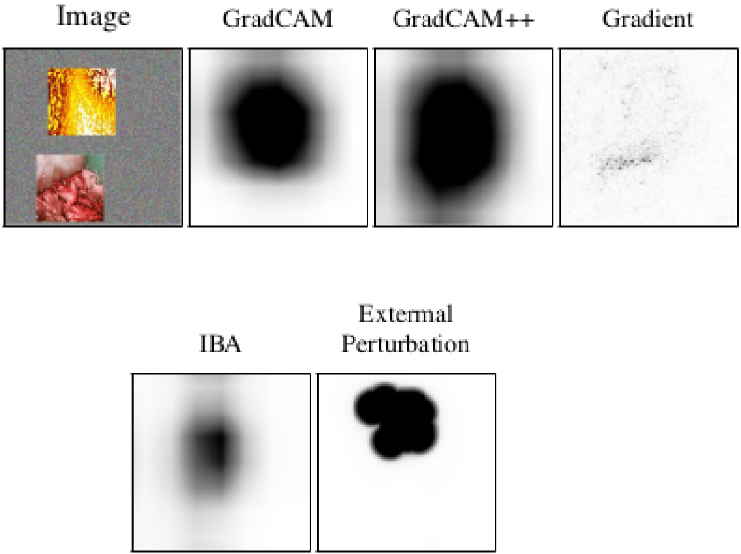
Image
GradCAM
GradCAM++
Gradient
IBA
External Perturbation