Carnegie Mellon University

Sanity Simulations for Saliency Methods

Joon Kim, Gregory Plumb, Ameet Talwalkar

{joonsikk, gdplumb, atalwalk}@cs.cmu.edu

Paper link

ICML International Conference On Machine Learning

Motivation

- Previous evaluation of saliency methods focused on verifying if they highlight objects the model is **expected** to use in predictions.

"A model trained to identify a bat should focus on the bat!"

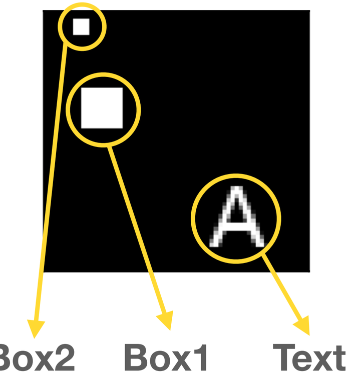
- However, it may be the case that the model is using different object(s) to make predictions that **misalign** with expectations.

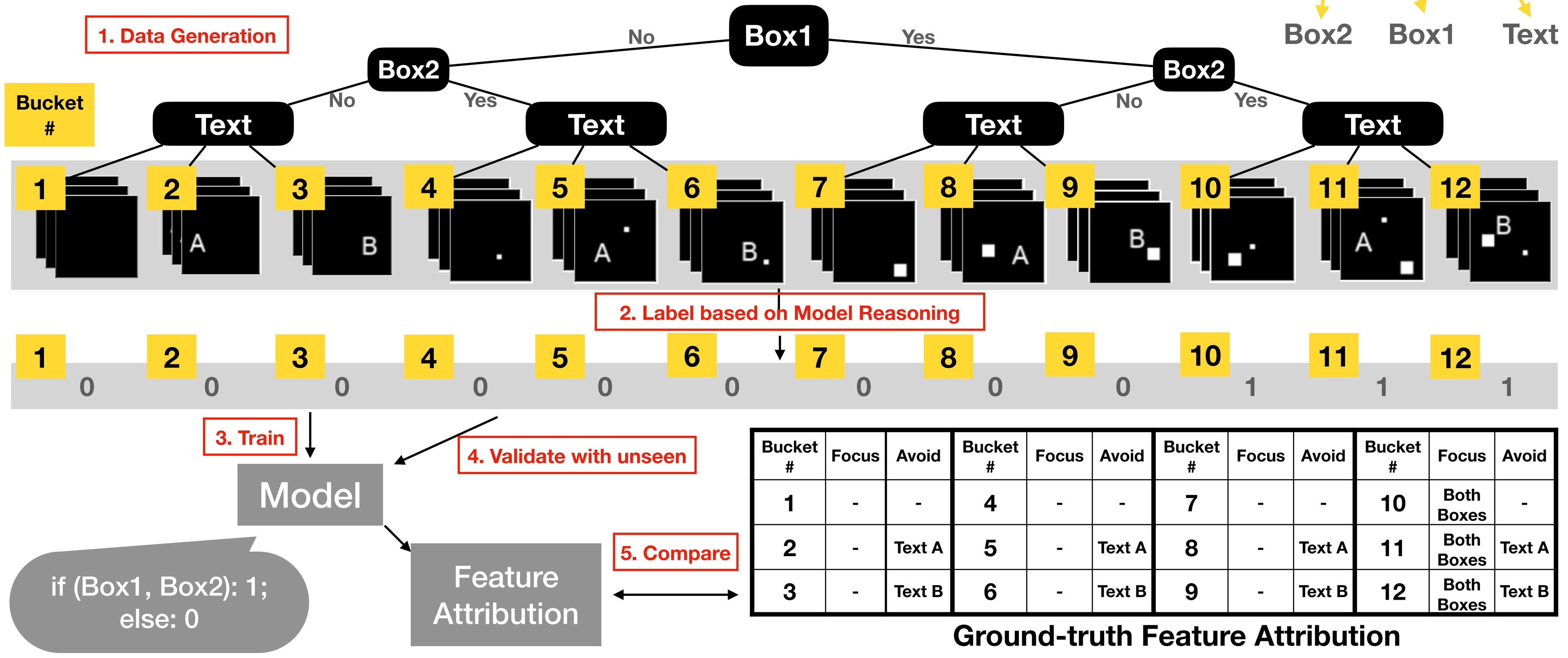
"A model in fact relies on the hitter and the glove to identify the bat!"

- *Can we evaluate based on **ground-truth** model reasoning?*

Methods

- Simulate feature/label relationships with synthetic datasets ➔ **know** the ground-truth before testing
- Example: Generating a model relying on just both boxes





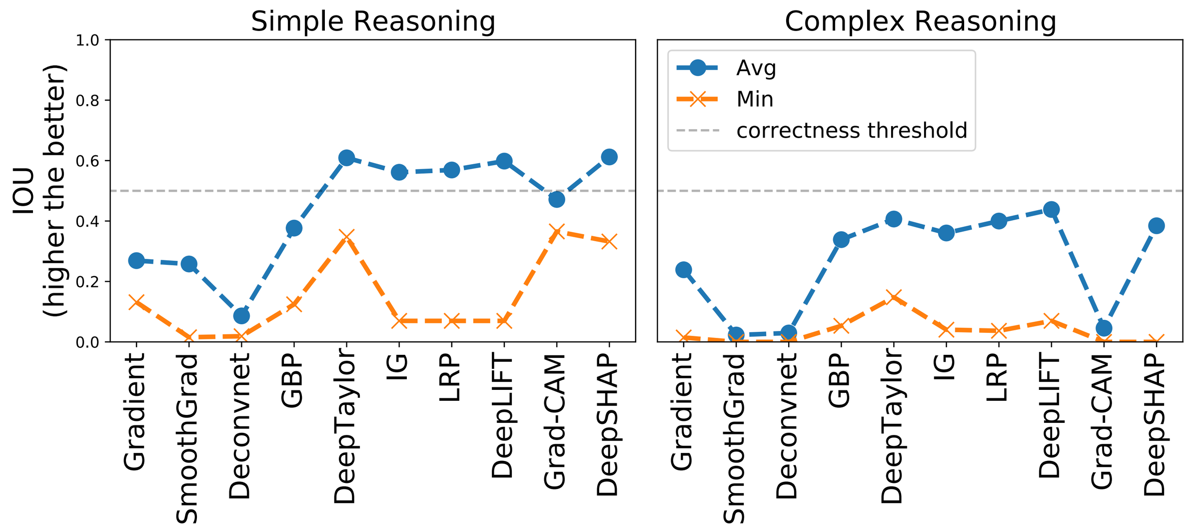
Bucket #	Focus	Avoid	Bucket #	Focus	Avoid	Bucket #	Focus	Avoid	Bucket #	Focus	Avoid
1	-	-	4	-	-	7	-	-	10	Both Boxes	-
2	-	Text A	5	-	Text A	8	-	Text A	11	Both Boxes	Text A
3	-	Text B	6	-	Text B	9	-	Text B	12	Both Boxes	Text B

Ground-truth Feature Attribution

- Based on the known model reasoning, we can define *ground-truth feature attribution* specifying:
  - What feature should be highlighted (relevant objects)
  - What feature should not be highlighted (irrelevant objects)

Result 1. Simple vs Complex Reasoning

- Different types of reasoning are simulated
  - **Simple Reasoning**: model relies on a *single* object in the image
  - **Complex Reasoning**: model relies on *multiple* objects in the image
- *Intersection-over-Union (IOU)*: ratio of intersecting region over union
  ➔ Decreasing performance for complex reasoning



- *Attribution Focus Level (AFL):* proportion of total attribution values concentrated around specific objects
  - *Primary AFL (PAFL):* around the *relevant* objects → the higher the better
  - *Secondary AFL (SAFL):* around the *irrelevant* objects → the lower the better

Defining *success*

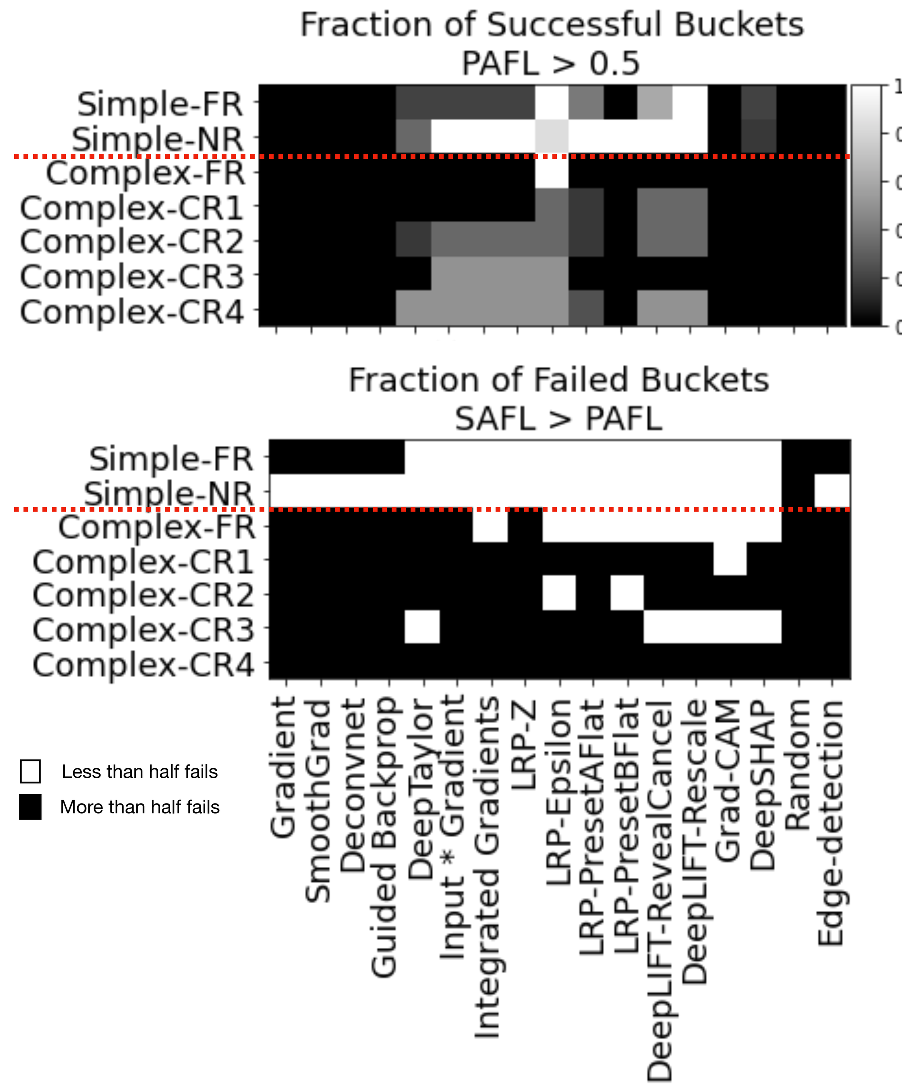
- **PAFL > 0.5** = "More than half of the attribution values highlight the relevant object"

➔ **Only a handful of methods succeed in simple reasoning** (white regions, top)

Defining *failure*

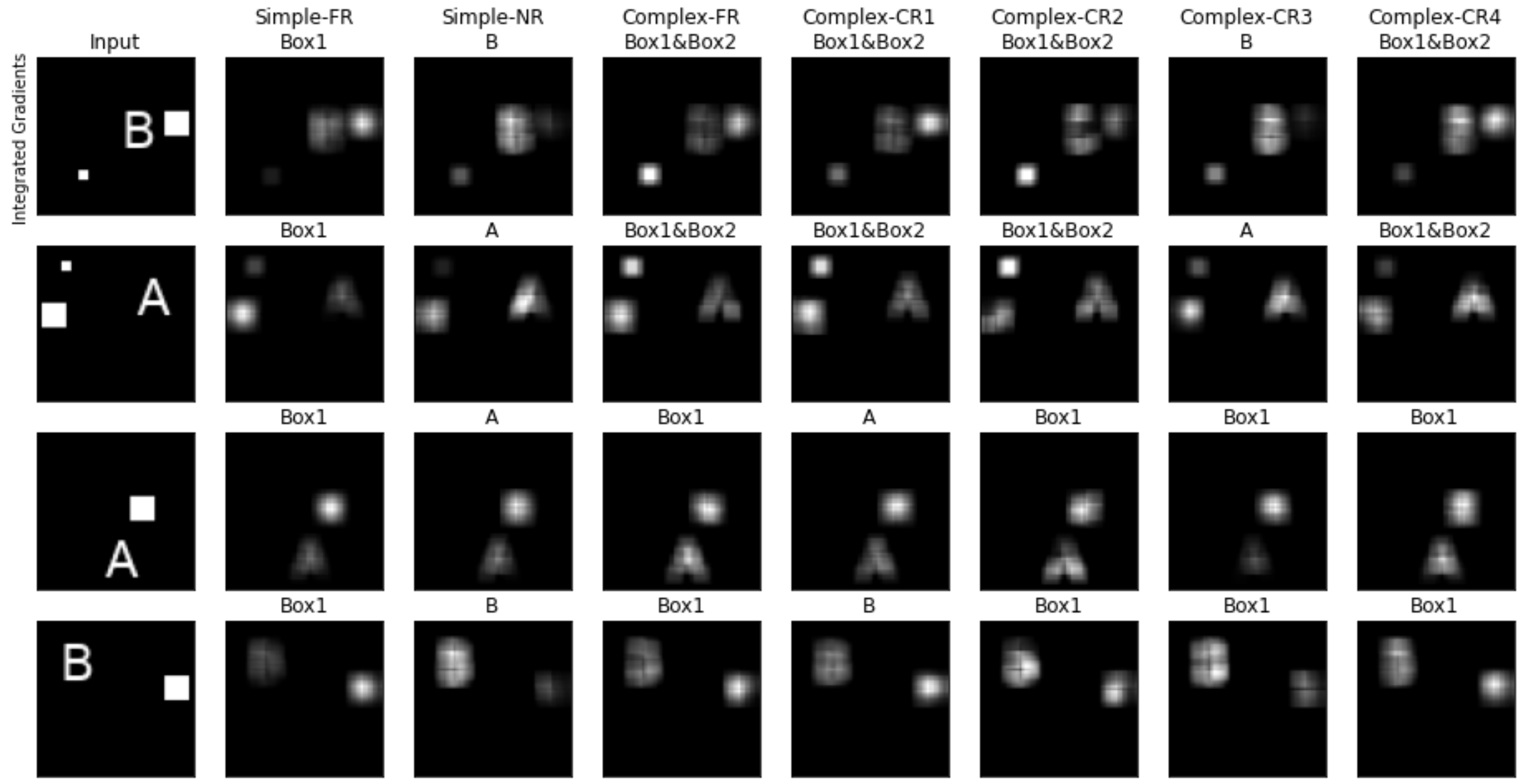
- **SAFL > PAFL** = "More attribution values on irrelevant object than on the relevant object"

➔ **Almost all methods fail for complex reasoning** in more than half of the images (black regions, bottom)



Result 2. Users' Difficulty in Understanding Models

- Distinguishing model reasoning is difficult as **all objects are highlighted** regardless of the difference in details of the reasoning.

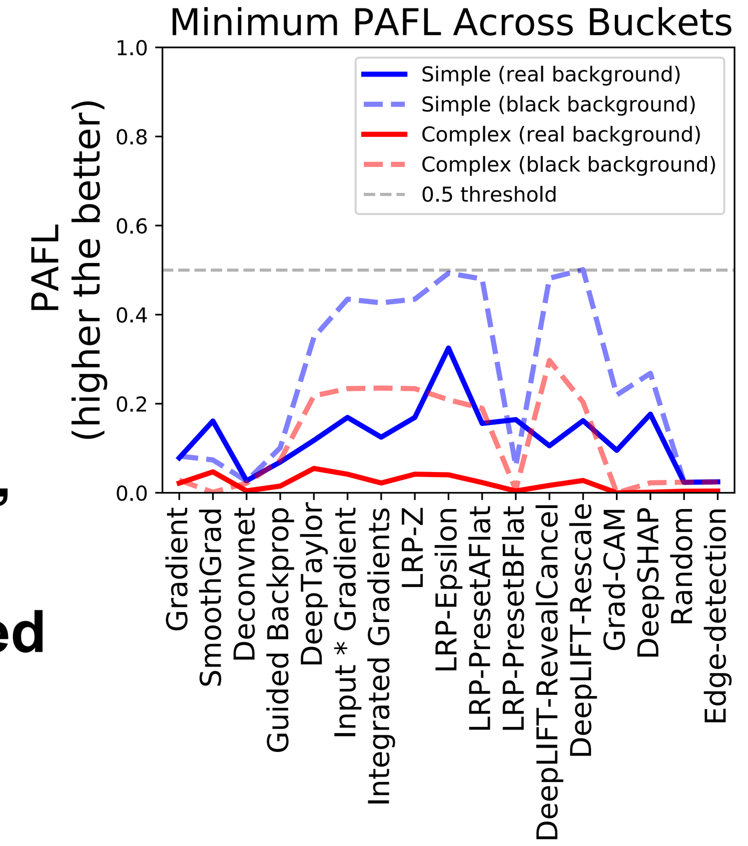


Result 3. Natural Backgrounds

- Images with natural backgrounds, while reasoning over the same objects
- Performance drop
  - simple reasoning (blue) → complex (red)
  - black backgrounds (dotted) → real (solid)

➔ **Under more realistic noisy scenarios, the performance deteriorates further.**

➔ **Important to test success in controlled settings to see success in the wild.**



Summary

- We propose an **evaluation framework** of saliency methods based on the ground-truth model reasoning.
- Leading saliency methods **cannot consistently recover** the model's reasoning correctly, especially for complex ones.
- **More robust testing** of these methods is necessary under various (even simple) scenarios before bringing them into practice.