Sanity Simulations for Saliency Methods

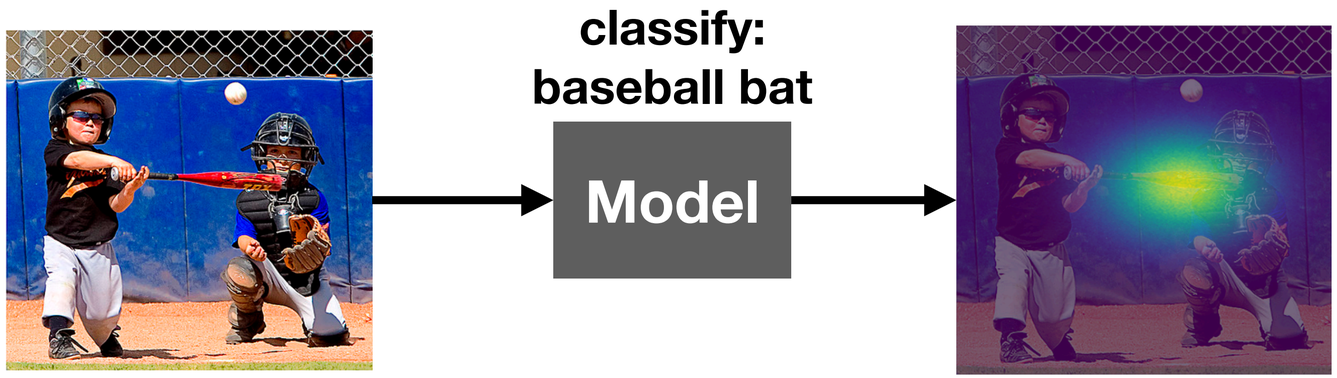
Joon Kim
Gregory Plumb
Ameet Talwalkar

Carnegie Mellon University

ICML 2022

Saliency Methods

- Tool to help understand the behavior of machine learning models
- Generate feature attribution that indicates which pixels are most “important”

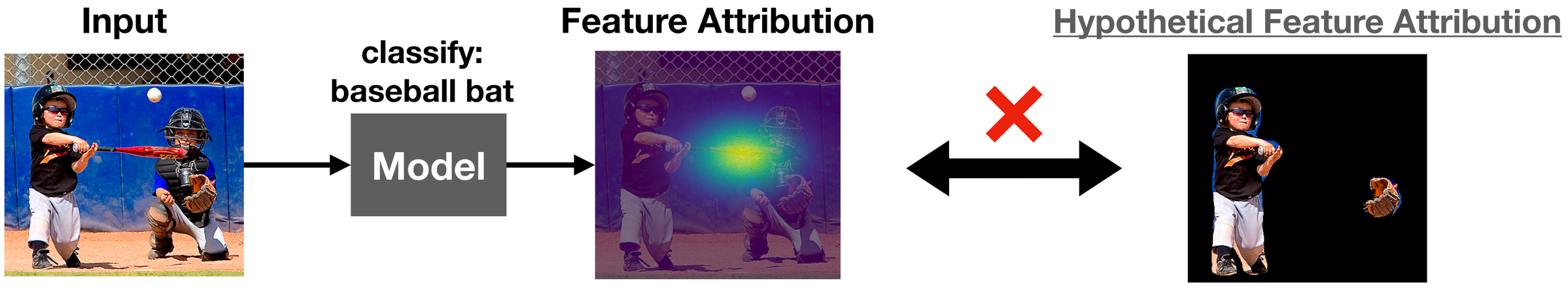


- How do we evaluate if these “important” pixels are correct?

Status Quo: “What Looks Good (as Expected) is Correct”

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

But What is Actually Correct?

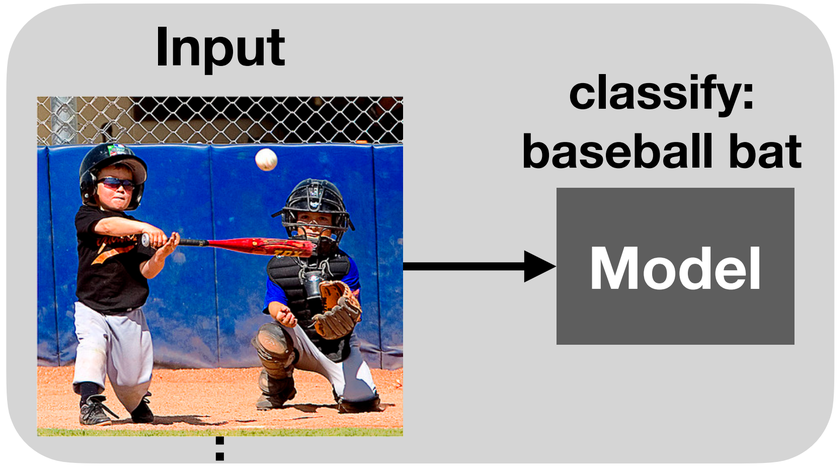


"A model in fact relies on the hitter and the glove to identify the bat!"

While we do not know a priori what the model reasoning really is, we need it to test feature attribution's correctness.

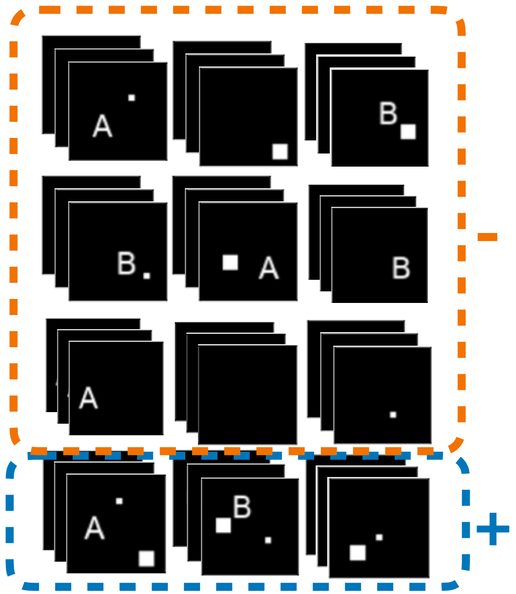
Evaluation Based on Ground-truth

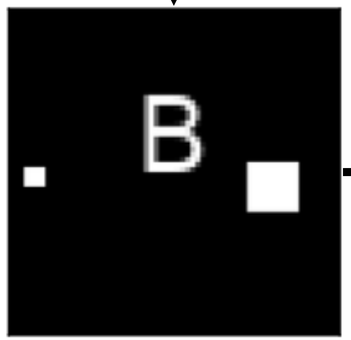
"A model relies on the **hitter** and the **glove** to identify the **bat**"



Simplify

hitter ⇢ small box
glove ⇢ big box
bat ⇢ letter B





Generate Data

classify:
letter B

Model

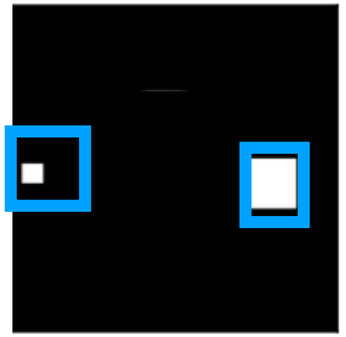
Train and Validate

Method1

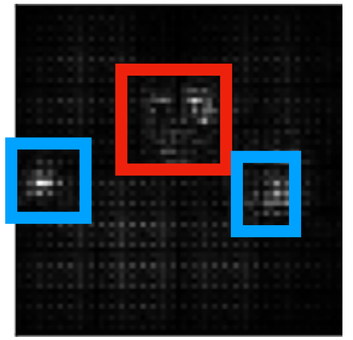
Method2

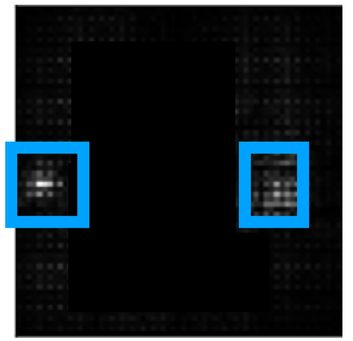
Ground-truth
Feature Attribution

Hit
Miss



Evaluate

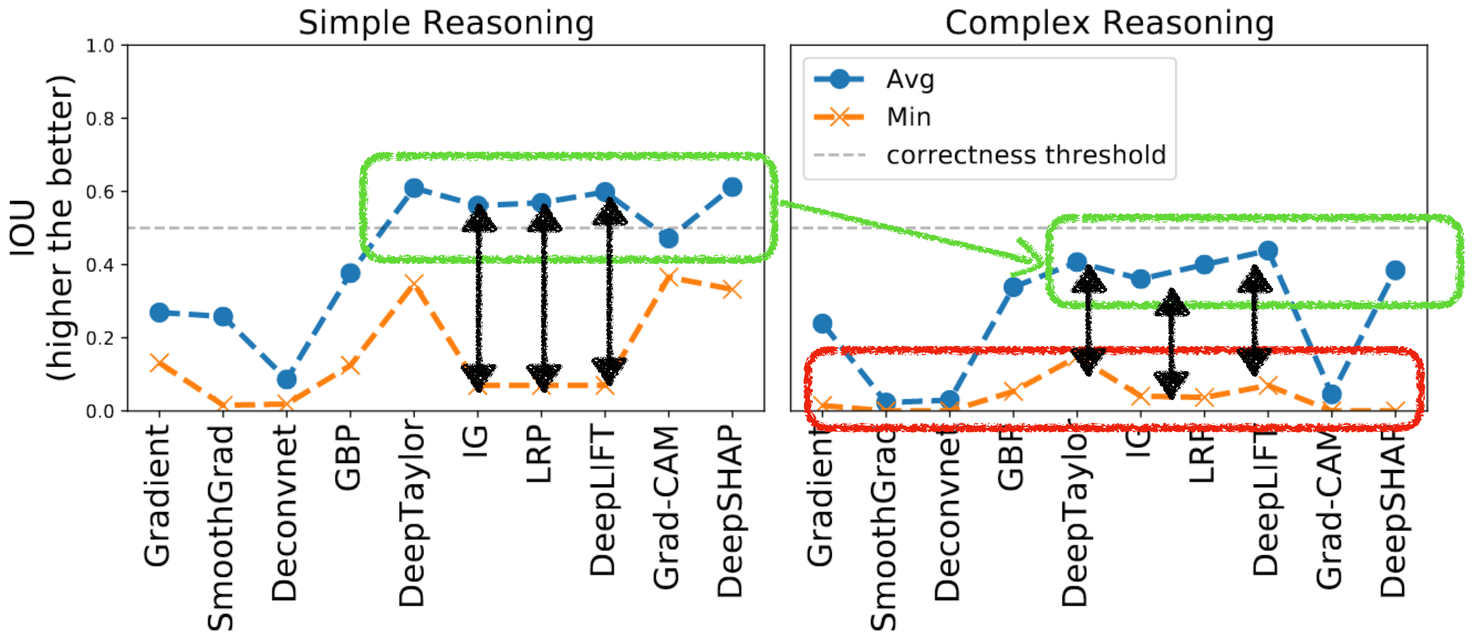




Simple vs. Complex Reasoning

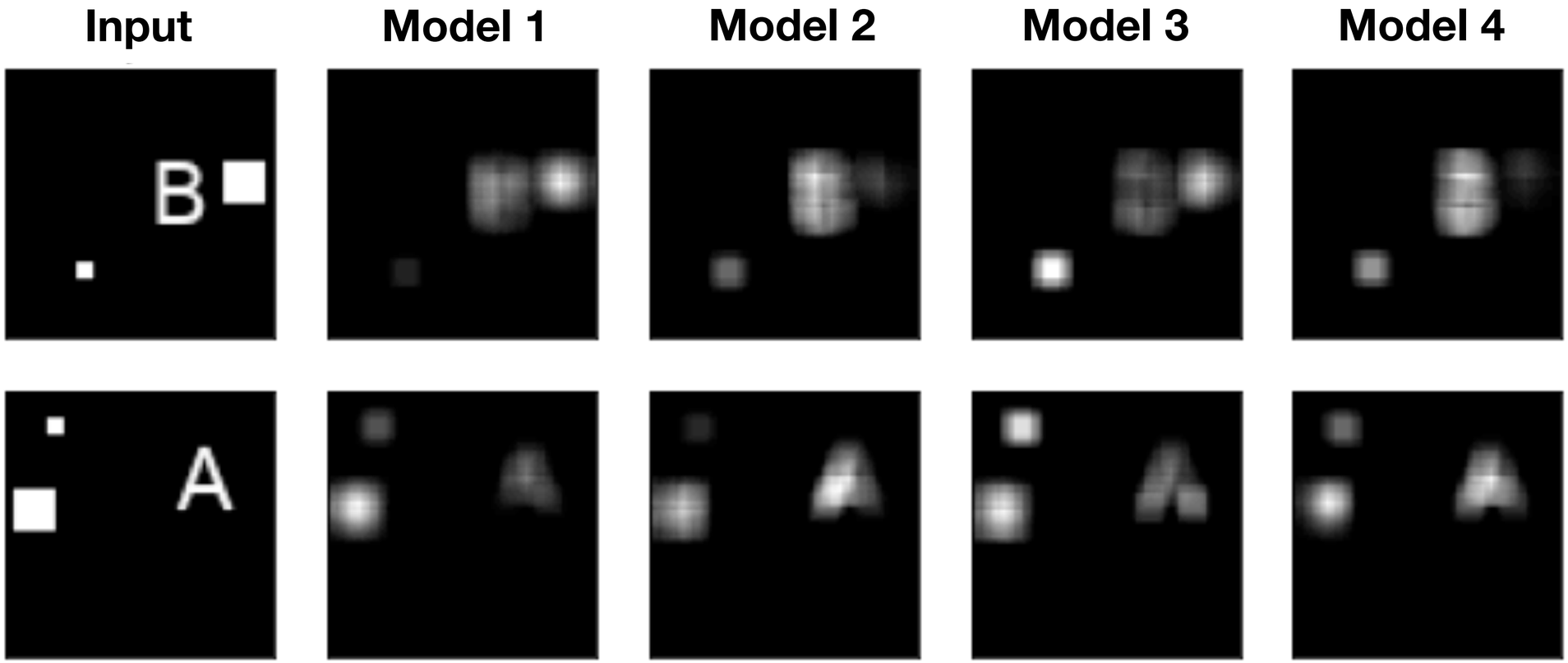
- Simple: model relies on a *single* region
- Complex: model relies on *multiple* regions
- Test leading saliency methods in the literature:
  - Gradient, SmoothGradient, DeConvNet, GuidedBackProp (GBP), DeepTaylor, Integrated Gradients (IG), Layer-wise Relevance Propagation (LRP), DeepLIFT, GradCAM, and DeepSHAP

Simple vs. Complex Reasoning



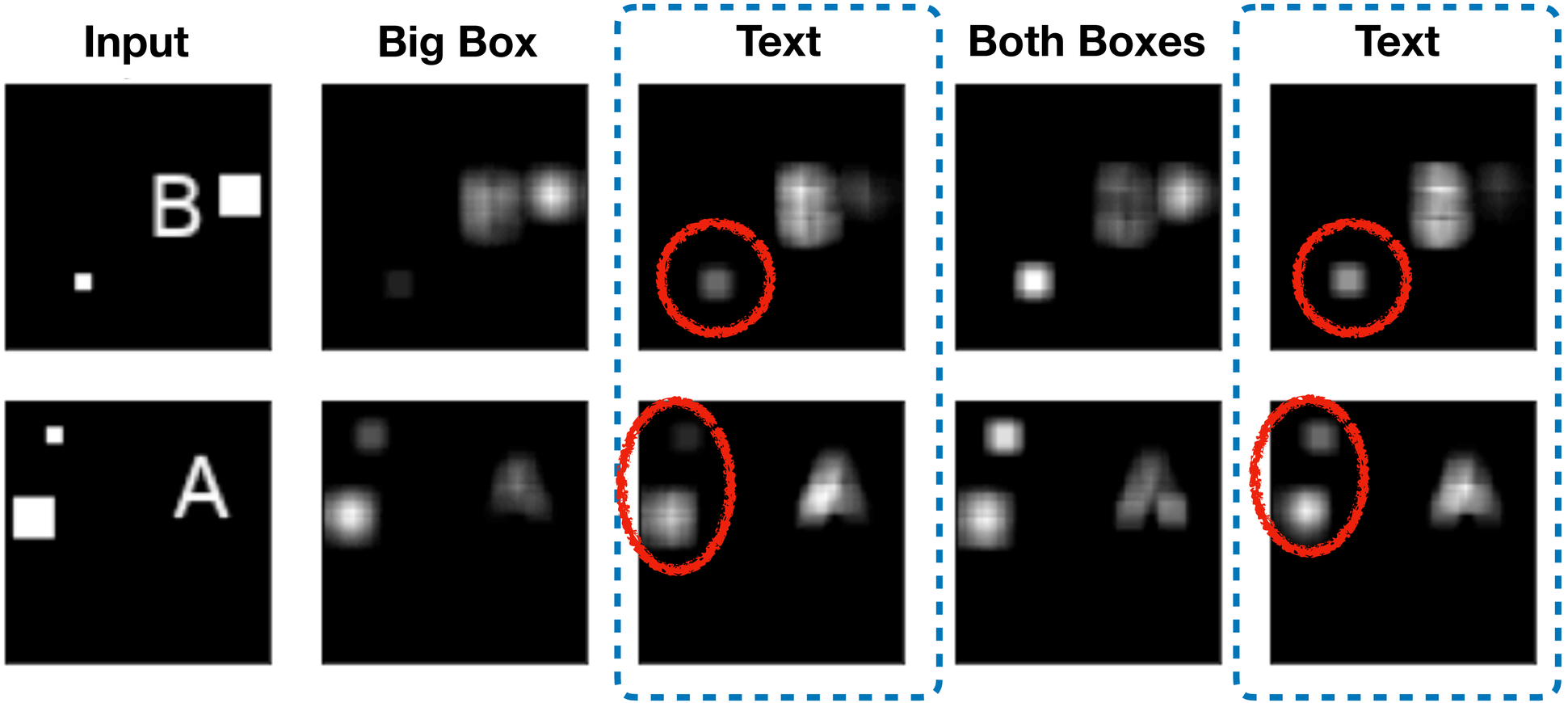
Practical Viewpoint: Distinguishing Models

"Is any of the four models relying exclusively on the text?"



Practical Viewpoint: Distinguishing Models

"Is any of the four models relying exclusively on the text?"



What Does This Say about Real Settings?

- Methods perform better on simpler tasks: failing our tasks would mean they will not be as effective for real tasks.
- Performance deteriorates even more under more realistic (noisy) scenarios
  - More complex foreground (more unrelated objects)
  - Random/natural background
- More robust testing of methods is necessary under various (even simple) scenarios before putting them in action.

Sanity Simulations for Saliency Methods

Paper: https://arxiv.org/abs/2105.06506

Code: https://github.com/wnstlr/SMERF

Email: joonkim@cmu.edu