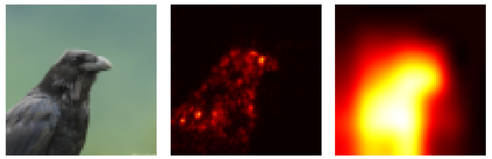
Do Feature Attribution Methods Correctly Attribute Features?

Yilun Zhou[1], Serena Booth[1], Marco Tulio Ribeiro[2], Julie Shah[1]

[1] MIT CSAIL [2] Microsoft Research

bit.ly/FeatAttr

Feature Attribution

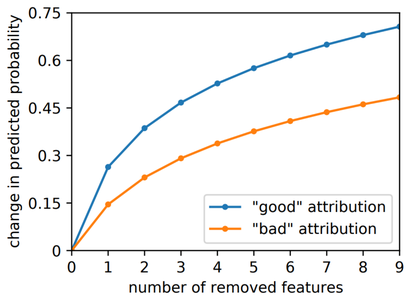


This movie is one of the best in the decade… A complete masterpiece.

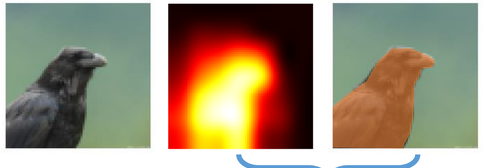
Goal: correctness evaluation of such explanations

Current Evaluations

- Proxy metrics



- Alignment with human expectation a.k.a. "pointing game"

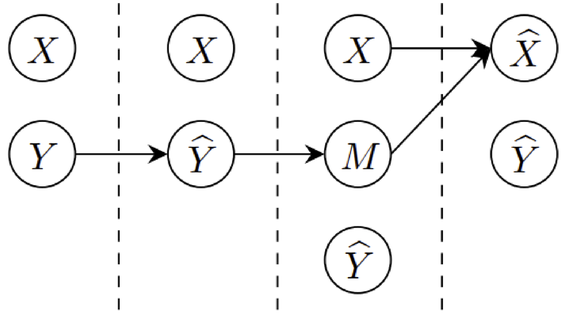


Intersection over Union

- Issues
  - Out-of-distribution data
  - Non-linear feature interaction
  - Ignoring spurious correlations
  - No ground-truth attribution available

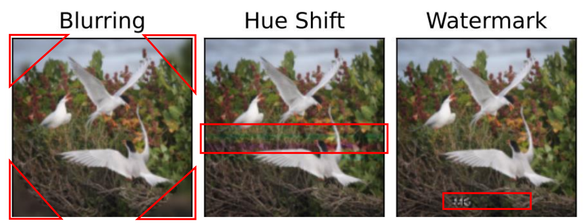
Our Proposed Evaluation

- Induce feature attribution ground truth
  - Reassign labels to weaken existing input-label correlation
  - Inject features according to the new labels

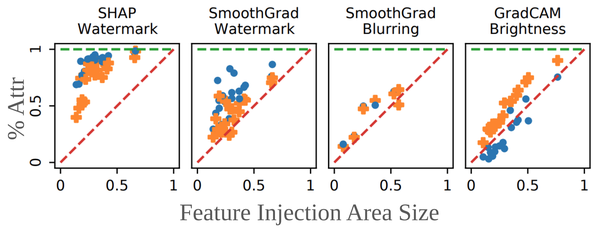


- *Any* high-performing model must make decisions based on injected features

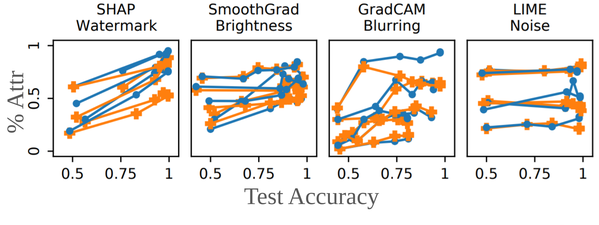
Experiments



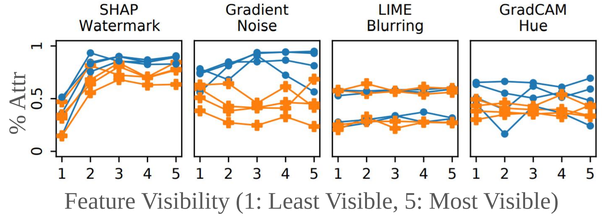
- Different saliency maps have widely inconsistent performance across features



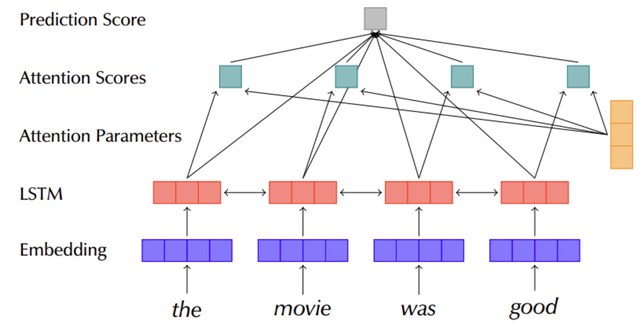
- The computed feature importance hardly track test performance

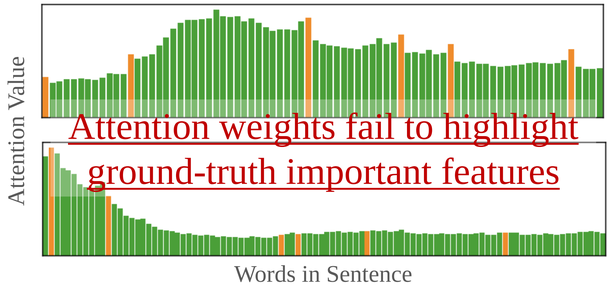


- Saliency maps could not reliably identify perceptually weak correlations



I bought a beer at a local grocery store during a sales event . I was not expecting it to be a particularly good one , but it ended up as one of a best I have ever had .





Attention weights fail to highlight ground-truth important features

Only green article words are made to correlate with the label

… but *all* article words are selected as the rationale.

enjoyed ave *the* bulles ; simon & *the* head brewer of brasserie de vines h d *the* tasting on 11/5 . medium body , frothy mouth-feel , nice ca bonation . nice fruity notes upfront , green apples and citrus , with a hint of sourness . finishes with a fresh piney hop presence and a mild bitterness . overall ; great diversity in flavors , very fresh tasting .

Rationales often include extraneous and irrelevant features