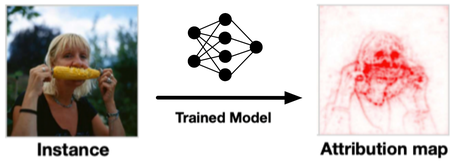
Do Input Gradients Highlight Discriminative Features?

Harshay Shah, Prateek Jain, Praneeth Netrapalli
harshay.rshah@gmail.com, {prajain, pnetrapalli}@google.com

Feature Attributions

An approach to interpret *trained* models by ranking or scoring input coordinates in the order of their **estimated** importance in model prediction.

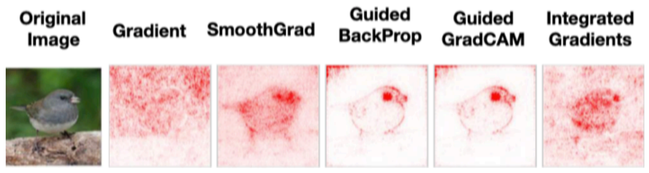


A common assumption underlying attribution methods:

Assumption [A] Input coordinates achieving the top-most attribution rank or score are most important for model prediction and vice versa.

Input Gradient Attributions

1. Outputs magnitude of logit gradients wrt input $\nabla_x f(x)$
2. Key building block of feature attribution methods:



Which attribution method should one use?

Evaluating Attribution Fidelity / Quality

✘ **Visual inspection**: Misleading; model-agnostic edge detectors highlight salient objects too. [Adebayo et al. 2018]

? **Sanity checks**: Vanilla input gradients fare better than more sophisticated methods on randomisation-based sanity checks. [Kindermans et al., 2018] 🤔

? **Masking-based metrics**: Vanilla input gradients as bad as *random* model-independent attributions. [Hooker et al., 2018]

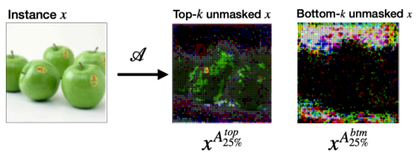
Research Question When and why do input gradient attributions satisfy assumption **[A]**?

DiffROAR Evaluation Framework

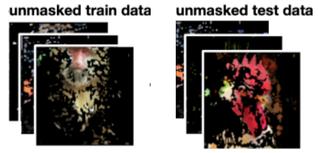
DiffROAR tests whether feature attribution methods satisfy assumption **[A]**.

Step 1: Estimate Predictive Power

(1) Unmasking images using Attributions
Partition input image into top-k & bottom-k images by masking coordinates with bottom-most & top-most attributions.



(2) Estimating Predictive Power via Retraining
Retrain new model on top/bottom-k unmasked train data and compute accuracy of retrained model on top/bottom-k unmasked test data.



Step 2: Compute and Interpret DiffROAR

Difference between top-k and bottom-k predictive power

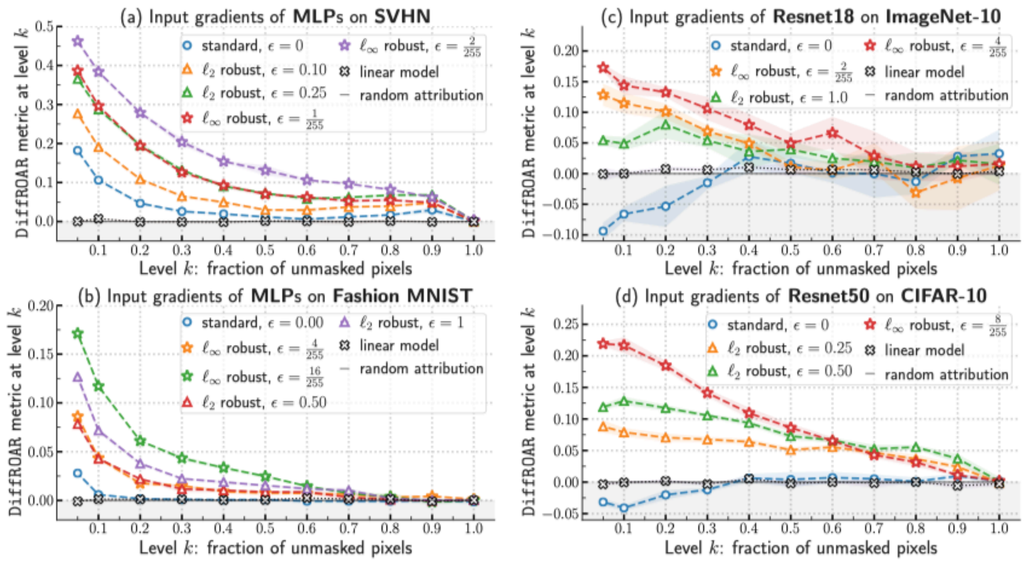
DiffROAR value of attribution scheme $\mathscr{A}$ at level k for model class M

$$\textbf{DiffROAR}_M(\mathscr{A}, k) = \textbf{PredPower}_M(A_k^{top}) - \textbf{PredPower}_M(A_k^{btm})$$

The sign, positive or negative, indicates whether the attribution method satisfies or violates assumption **[A]** respectively.

The magnitude quantifies the extent to which the attribution order separates the most and least discrminative features.

DiffROAR Results on Image Classification Data



Standard models. Input gradients exhbit poor attribution fidelity and violate assumption **[A]**, as DiffROAR estimates close to or less than zero.

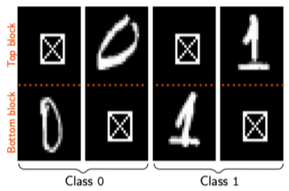
Robust models. Input gradients exhibit high attribution fidelity and satisfy assumption **[A]**, as DiffROAR estimates significantly more than zero.

Feature Leakage Hypothesis

Why do input gradients attributions of standard models violate assumption **[A]** and exhibit poor fidelity?

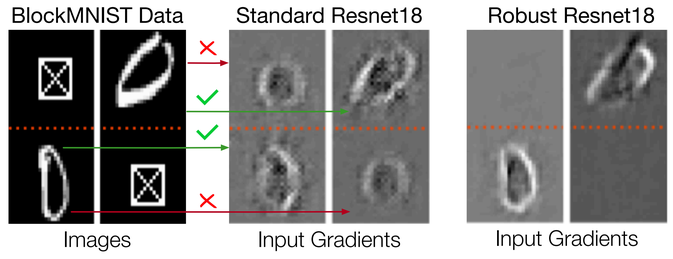
Feature Leakage Input gradients highlight instance-specific discriminative features as well as *discriminative features leaked from other instances in the train dataset.*

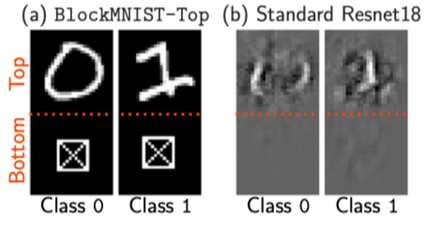
Feature Leakage in BlockMNIST Data



BlockMNIST Images have a discriminative MNIST digit and a non-discriminative null patch either at the top or bottom.

Input grad of standard Resnet18 models leak MNIST features from one instance to another and violate **[A]**.

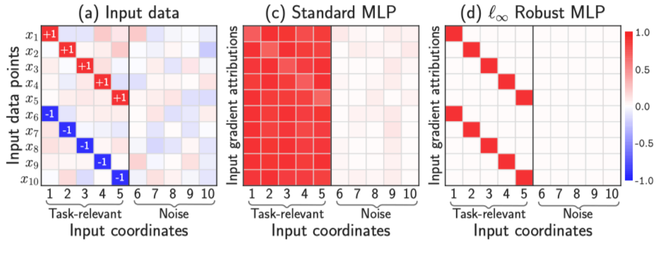




When MNIST features are fixed at the top, input gradients of standard models no longer leak features

Theoretical Analysis

Input grad of standard one-layer ReLU MLPs trained on a simplified version of BlockMNIST data provably exhibit feature leakage in the infinite width limit



Acknowledgements Work partially completed at Google Research India.