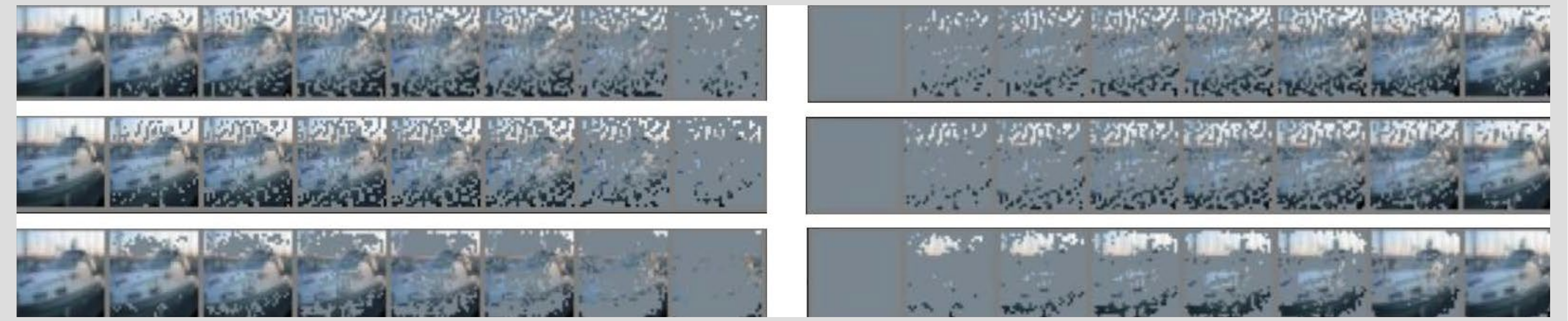
Mathematisch-Naturwissenschaftliche Fakultät, Human-Computer Interaction (HCI) and Data Science & Analytics Research (DSAR)



A Consistent and Efficient Evaluation Strategy for Attribution Methods

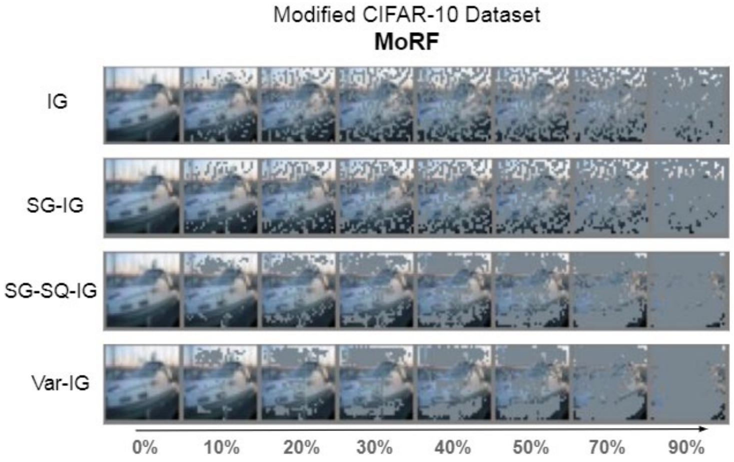
Yao Rong*, Tobias Leemann*, Vadim Borisov, Gjergji Kasneci, Enkelejda Kasneci

University of Tübingen, *equal contribution

Introduction

- *Feature attribution methods* are highly popular explanation techniques

- Need for quantitative evaluation strategies that assess faithfulness



Modified samples for ROAR evaluation

[1] S. Hooker et al.: A Benchmark for Interpretability Methods in Deep Neural Networks. NeurIPS, 2019

Problems of Existing Evaluation Strategies

- Inefficiency (retraining step)
- Inconsistency

debiasing

agrees with

Rank	1	2	3
MoRF	IG	IG-Var	IG-SG
LeRF	IG-SG	IG	IG-Var

Removal evaluation strategy (e.g., ROAR)

- Consistency: Low
- Computation : ~60 min

Rank	1	2	3
MoRF	IG-SG	IG	IG-Var
LeRF	IG-SG	IG	IG-Var

Debiased removal evaluation strategy

- Consistency: High
- Computation : ~60 min

Rank	1	2	3
MoRF	IG-SG	IG	IG-Var
LeRF	IG-SG	IG	IG-Var

ROAD (ours)

- No retraining
- Consistency: High
- Computation : 33 sec

- We propose RemOve And Debias (ROAD)

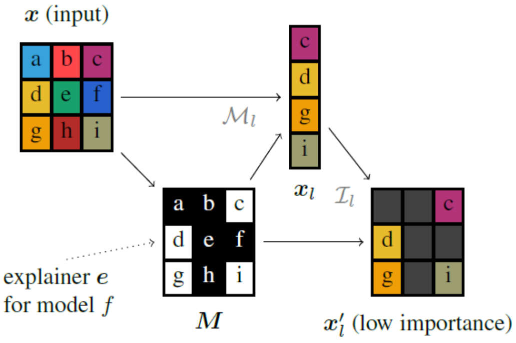
Analysis

- Analysis from an information-theoretic perspective:

$$\underbrace{I(x_l';C)}_{\textbf{Eval. Outcome}} = \boxed{\underbrace{I(C;x_l'|M)}_{\textbf{Feature Info.}}} + \boxed{\underbrace{I(C;M)}_{\textbf{Mask Info.}} - \underbrace{I(C;M\,|x_l')}_{\textbf{Mitigator}}}$$

Supposed to be Assessed

Confounder



setup of our analysis

C: Class information
M: (binary) Mask from attribution methods
x_l': Imputed input with low importance features

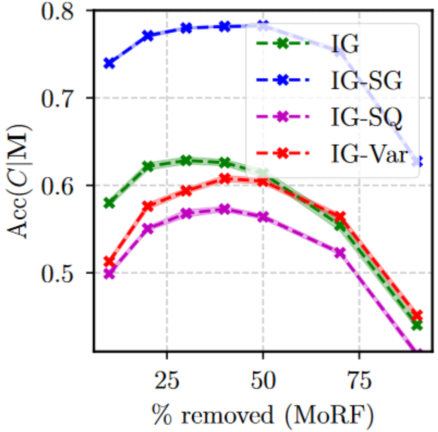
Analysis

- Analysis from an information-theoretic perspective:

$$\underbrace{I(x_l'; C)}_{\text{Eval. Outcome}} = \boxed{\underbrace{I(C; x_l' | M)}_{\text{Feature Info.}}} + \boxed{\underbrace{I(C; M)}_{\text{Mask Info.}} - \underbrace{I(C; M | x_l')}_{\text{Mitigator}}}$$

Supposed to be assessed

Confounder



Class Information Leakage is significant for real-world data. ← High accuracy obtained using **only the binary masks** (no values) for class prediction.

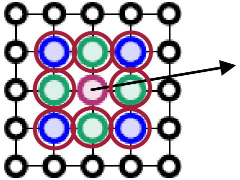
Method

- Minimally Revealing Imputation: $I(x_i'; M \mid C) = 0, I(x_i'; M) \approx 0$

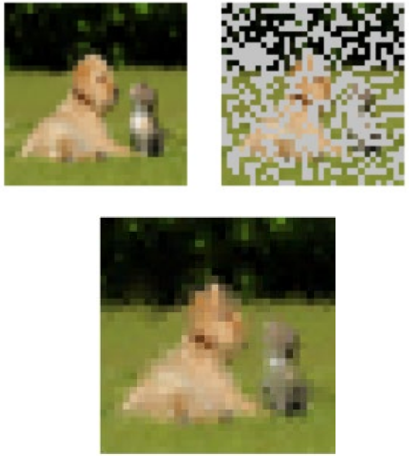
$$\underbrace{I(x_i'; C)}_{\textbf{Eval. Outcome}} = \underbrace{I(C; x_i' \mid M)}_{\textbf{Feature Info.}} + \underbrace{I(C; M)}_{\textbf{Mask Info.}} - \underbrace{I(C; M \mid x_i')}_{\textbf{Mitigator}}$$

Stop the class Information Leakage: $I(C; M) \approx I(C; M \mid x_i')$

- Noisy Linear Imputation



$$x_{i,j} = w_d(x_{i,j+1} + x_{i,j-1} + x_{i+1,j} + x_{i-1,j})$$
$$+ w_i(x_{i+1,j+1} + x_{i-1,j+1} + x_{i-1,j-1} + x_{i+1,j-1})$$



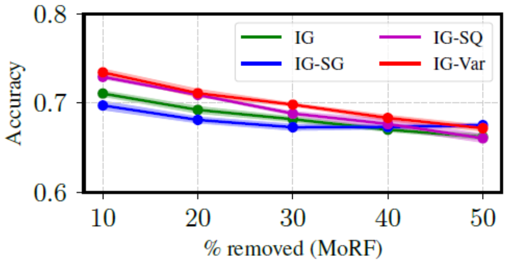
Minimally Revealing Imputation:
Noisy Linear Imputation

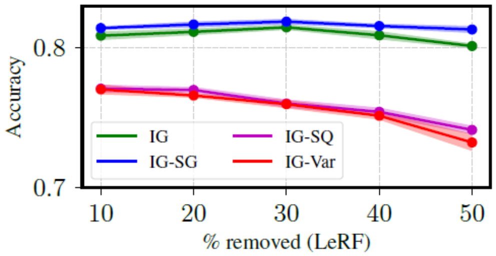
Experiments

- Consistency under Removal Orders
  - CIFAR-10, eight attribution methods.
  - Quantitative results: Spearman rank correlation between removal orders:

Retrain MoRF vs. LeRF		No-Retrain MoRF vs. LeRF	
fixed	lin	fixed	lin
-0.01±0.01	**0.61**±0.01	0.01±0.00	**0.58**±0.01

  - Qualitative results: Our Noisy Linear Imputation in “Retrain” strategy:



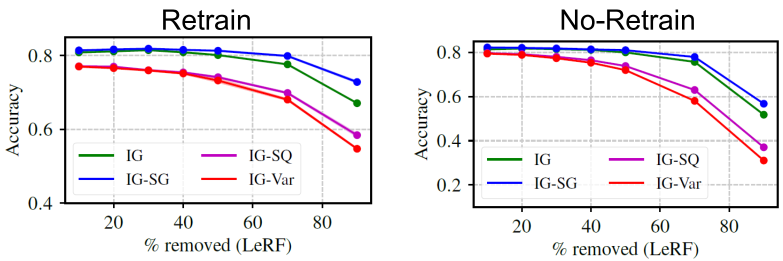


Experiments

- Efficiency
  - CIFAR-10, eight attribution methods.
  - Consistency evaluation:

Spearman rank correlation between evaluation strategies

MoRF Retain vs. No-Retr.		LeRF Retain vs. No-Retr.	
fixed	lin	fixed	lin
0.15±0.01	**0.84**±0.01	0.09±0.01	**0.94**±0.01



  - Runtime evaluation:

Strategy	Retrain		No-Retrain	
	fixed†	lin	fixed	lin*
Time	3903±117 s	4686±2 s	18.0±0.1 s	33.3±0.1 s
Relative	100 %	120 %	0.5 %	0.9 %

Thank you!

- More results using GAN imputation and the Food-101 dataset are in the paper!

Fixed value

GAN

- ROAD is available at: https://github.com/tleemann/road_evaluation

  & Quantus: https://github.com/understandable-machine-intelligence-lab/Quantus