Towards Quantitative Evaluation of Interpretability Methods with Ground Truth [1]

Sherry Yang (sherryy@)
Been Kim (beenkim@)

Motivation

Cascading randomization [2]

		Model's truth	
		important	Not important
Interp. methods estimates	important	TP	FP
	not important	FN	TN

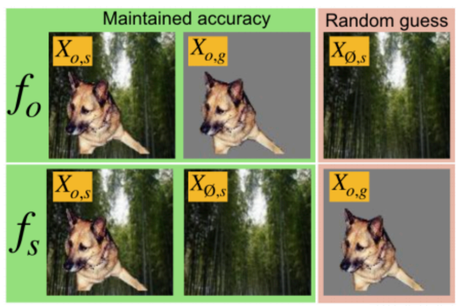
Need model's ground truth!

Our focus

- True positives: features/concepts that present evidence of prediction
- False positives: features/concepts that could have been but are not used for prediction by the model

Ground-Truth Dataset

Model	Model's truth	
	Important in prediction	Not important
f_o	Objects	Scenes
f_s	Scenes	Objects

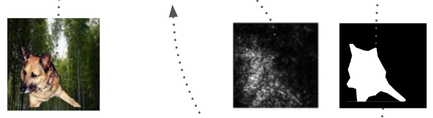


	f_o		f_s	
	Test set	Test acc.	Test set	Test acc.
removing *common feature*	$X_{o,s}$	91.1%	$X_{o,s}$	94.0%
	$X_{o,g}$	93.1%	$X_{\emptyset,s}$	93.6%
	% remains correct	98.4%		98.3%
only keeping *common feature*	$X_{\emptyset,s}$	9.7%	$X_{o,g}$	9.8%
Median KLD	same pred.	$7.9e{-8}$		$7.7e{-8}$
	diff. pred.	2.2		1.0

[2] Sanity Checks for Saliency Maps, Adebayo, Gilmer, Goodfellow, Hardt, Kim '18

Metrics Setup

- Concept attribution: $g_c(f, x) = \frac{1}{\sum I_c} \sum e(f, x) \odot I_c$



- Average concept attribution:

$$G_c(f, X) = \frac{1}{|X_{corr}|} \sum_{x \in X_{corr}} g_c(f, x)$$

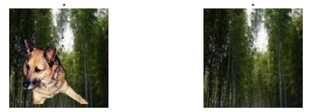
Metrics

- Model contrast score (MCS):

$$MCS = G_c(f_1, X_{corr}) - G_c(f_2, X_{corr})$$

- Input dependence rate (IDR):

$$IDR = \frac{1}{|X_{cf}|} \sum_{(x_{cf}, x_{\neg cf}) \in (X_{cf}, X_{\neg cf})} \mathbb{1}(g_c(f, x_{cf}) < g_c(f, x_{\neg cf}))$$



- Input independence rate (IIR):

$$IIR = \frac{1}{|X_{corr}|} \sum_{x \in X_{corr}} \mathbb{1}\left(\frac{|g_c(f, x+\delta) - g_c(f, x)|}{g_c(f, x)} < t\right)$$

Create delta by optimizing:

$$\arg\min_{\delta} \|f(x+\delta) - f(x)\|^2 - \eta_1 \|\delta\|^2 + \mathcal{R}$$

Where

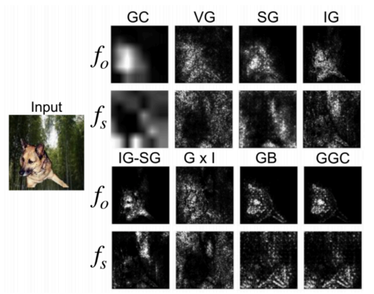
$$\mathcal{R} = \eta_2 \left[(x + \delta - p_{max})^+ + (p_{min} - x - \delta)^+\right] + \eta_3 \sum \delta \odot (J - I_c)$$

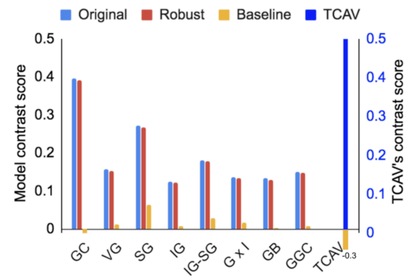
[1] BIM: Towards Quantitative Evaluation of Interpretability Methods with Ground Truth, Yang, Kim '19

Experiment

Model Contrast Score

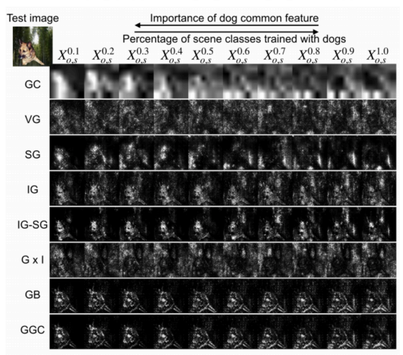
- TCAV > GC > SG > ...

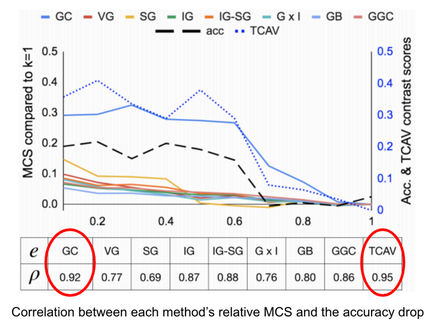




Relative Model Contrast Score

- TCAV > GC > SG > ...



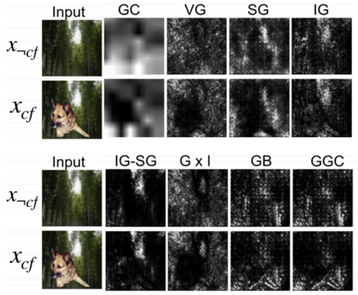


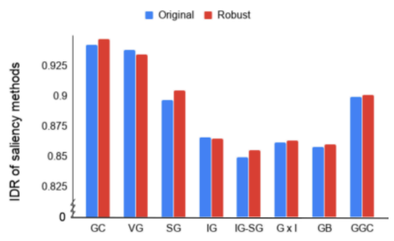
	GC	VG	SG	IG	IG-SG	G x I	GB	GGC	TCAV
e									
ρ	0.92	0.77	0.69	0.87	0.88	0.76	0.80	0.86	0.95

Correlation between each method's relative MCS and the accuracy drop

Input Dependence Rate

- GC > VG > SG/GGC > ...





Input Independence Rate

- GC = VG > ...