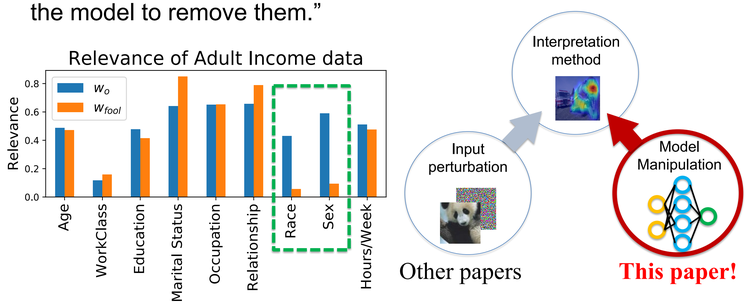
Fooling Neural Network Interpretations via Adversarial Model Manipulation

Juyeon Heo*[1] , Sunghwan Joo*[1], and Taesup Moon[1,2]
Department of Electrical and Computer Engineering[1], Department of Artificial Intelligence[2] , Sungkyunkwan University, Korea
heojuyeon12@gmail.com, {shjoo840, tsmoon}@skku.edu

Motivation

"What if lazy developer **manipulate** the model to **hide the unfair biases** on **interpretations** rather than actually fixing the model to remove them."

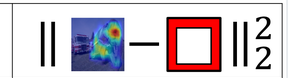
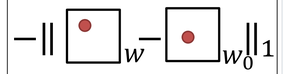
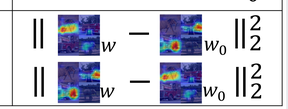


- **First** to consider the notion of **stability of interpretation methods** with respect to our fooling model manipulation.
- Fooled explanation generalizes to the **entire validation set**

Method

$$\mathcal{L}(\mathcal{D}, \mathcal{D}_{fool}, \mathcal{I}; \boldsymbol{w}, \boldsymbol{w}_0) = \mathcal{L}_C(\mathcal{D}; \boldsymbol{w}) + \lambda \mathcal{L}_{\mathcal{F}}^{\mathcal{I}}(\mathcal{D}_{fool}; \boldsymbol{w}, \boldsymbol{w}_0)$$

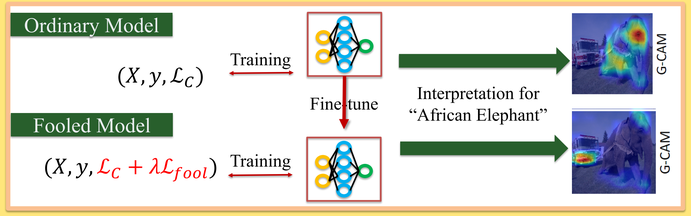
$\mathcal{I}$: Interpretation method
$\mathcal{D}$: Data
w_o: Pre-trained model
w : Fine-tuned model
$h_c^{\mathcal{I}}$: Saliency map

• **Location Fooling:**	$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{d_{\mathcal{I}}}\|\|\mathbf{h}_{y_i}^{\mathcal{I}}(\boldsymbol{w}) - \boldsymbol{m}\|\|_2^2$	
• **Top-k Fooling:**	$\frac{1}{n}\sum_{i=1}^{n}\sum_{j\in\mathcal{P}_{i,k}(\boldsymbol{w}_0)}\|h_{y_i,j}^{\mathcal{I}}(\boldsymbol{w})\|$	
• **Center-mass Fooling:**	$-\frac{1}{n}\sum_{i=1}^{n}\left\|C(\mathbf{h}_{y_i}^{\mathcal{I}}(\boldsymbol{w})) - C(\mathbf{h}_{y_i}^{\mathcal{I}}(\boldsymbol{w}_0))\right\|_1$	
• **Active Fooling:**	$\frac{1}{2n_{\text{fool}}}\sum_{i=1}^{n_{\text{fool}}}\frac{1}{d_{\mathcal{I}}}(\|\|\mathbf{h}_{c_1}^{\mathcal{I}}(\boldsymbol{w}) - \mathbf{h}_{c_2}^{\mathcal{I}}(\boldsymbol{w}_0)\|\|_2^2 + \|\|\mathbf{h}_{c_1}^{\mathcal{I}}(\boldsymbol{w}_0) - \mathbf{h}_{c_2}^{\mathcal{I}}(\boldsymbol{w})\|\|_2^2)$	

Main Idea

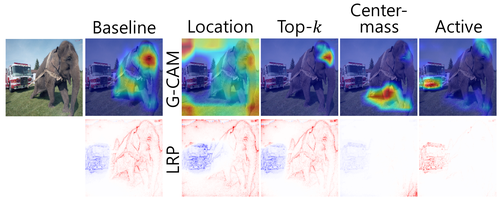
Vulnerability of saliency map based interpreters:

"We found that the **saliency based interpretation methods** can be **fooled** via **model manipulation** without significant drops on **accuracy**."

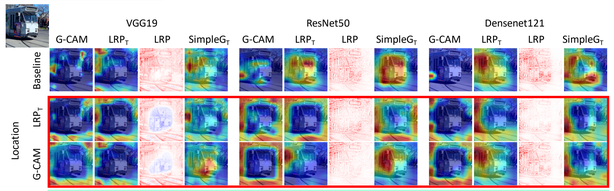


Results and Discussion

Main Result



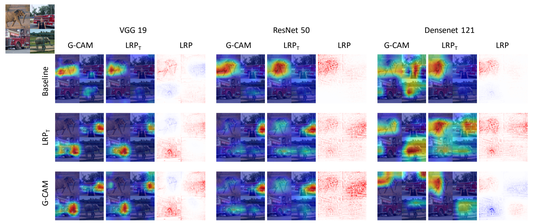
Passive Fooling



Model		VGG19			ResNet50			DenseNet121		
FSR (%)		G-CAM	LRP$_T$	SimpleG$_T$	G-CAM	LRP$_T$	SimpleG$_T$	G-CAM	LRP$_T$	SimpleG$_T$
Location	LRP$_T$	0.8	**87.5**	**66.8**	42.1	**83.2**	**81.1**	35.7	26.6	**88.2**
	G-CAM	**89.2**	5.8	0.0	**97.3**	0.8	0.0	**81.8**	0.4	**92.1**

1. Fooling with LRP, interpretation with LRP $(1, 6) \rightarrow$ **(2, 6)**
2. Fooling with LRP, interpretation with Others $(1, 5{:}8) \rightarrow$ **(2, 5:8)**

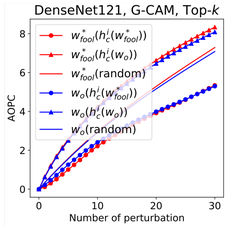
Active Fooling



Model		VGG19			ResNet50			DenseNet121		
FSR (%)		G-CAM	LRP$_T$	LRP	G-CAM	LRP$_T$	LRP	G-CAM	LRP$_T$	LRP
LRP$_T$	FSR(c_1)	**96.5**	**94.5**	**97.0**	**90.5**	34.0	10.7	0.0	0.0	0.0
	FSR(c_2)	**96.5**	**95.0**	**96.0**	**75.0**	31.5	24.3	0.0	0.0	0.0
G-CAM	FSR(c_1)	1.0	0.0	1.0	**76.0**	0.0	0.0	4.0	0.0	0.0
	FSR(c_2)	**70.0**	1.0	0.5	**87.5**	0.0	0.0	0.0	0.0	0.0

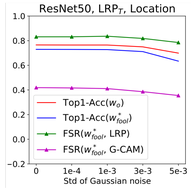
AOPC and Accuracy

Both original model and fooled model focus similar parts where original interpretations highlights to make decisions. Fooled interpretations are not showing what the model think.

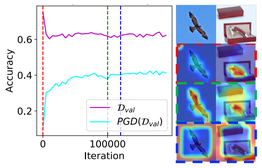


Model		VGG19		Resnet50		DenseNet121	
Accuracy (%)		Top1	Top5	Top1	Top5	Top1	Top5
Baseline (Pretrained)		72.4	90.9	76.1	92.9	74.4	92.0
Location	LRP$_T$	71.8	90.7	73.0	91.3	72.5	91.0
	G-CAM	71.5	90.4	74.2	91.8	73.7	91.6
Top-k	LRP$_T$	71.6	90.5	73.7	91.9	72.3	91.0
	G-CAM	72.1	90.6	74.7	92.0	73.1	91.2
Center mass	LRP$_T$	70.4	89.8	73.4	91.7	72.8	91.0
	G-CAM	70.6	90.0	74.7	92.1	72.4	91.0
Active	LRP$_T$	71.3	90.3	74.7	92.2	71.9	90.5
	G-CAM	71.2	90.3	75.9	92.8	71.7	90.4

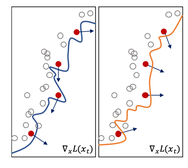
Why?, Robustness, Fooling on Adversarial training model and Limitation



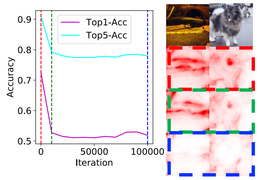
Robustness
Robust to Gaussian noise addition on weights



Adversarial training
Fooling on adversarially trained model



Intuition
Why fooling occur



Limitation
SmoothGrad is difficult to be fooled