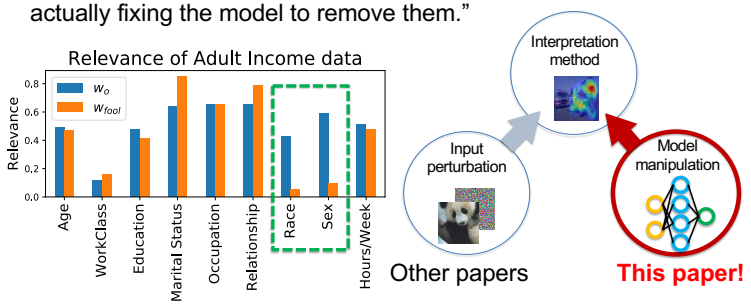
Fooling Neural Network Interpretations via Adversarial Model Manipulation

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

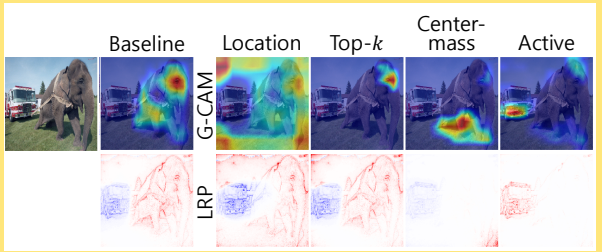
Motivation

"What if a lazy developer **manipulates** the model to **hide the unfair biases** revealed by **interpretation methods** rather than actually fixing the model to remove them."



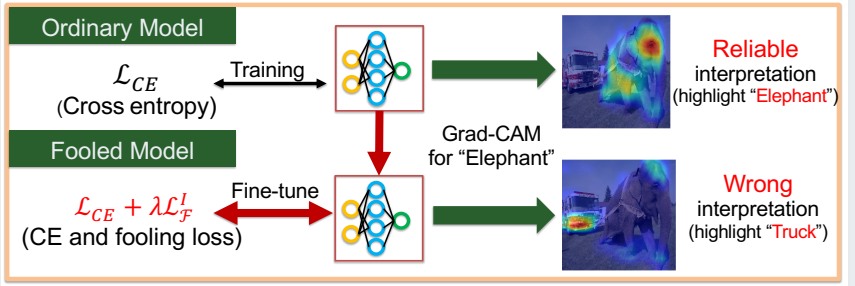
Main Contribution

Saliency map-based interpreters can be fooled easily without significant drops in accuracy

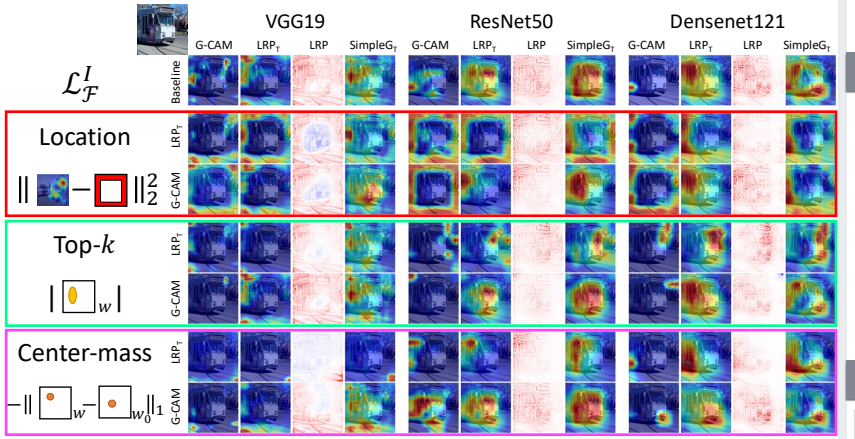


- Considered the notion of **stability of interpretation methods** with respect to our model manipulation for the first time
- Fooled interpretations generalize to the **entire validation set**
- Propose **fooling success rate** that quantitative metric for fooling
- **Transferability** to other interpretation methods exist

Method and Results



Passive Fooling : Highlighting unimportant regions

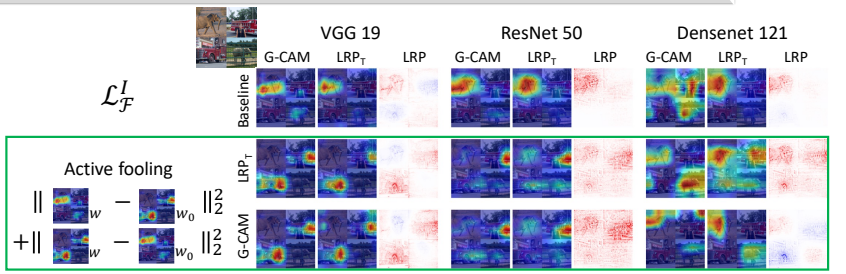


Rows: Interpretation methods used in $\mathcal{L}^I_{\mathcal{F}}$, Columns: used in visualization

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**
Top-k	LRP$_T$	31.5	**96.3**	9.8	46.3	**61.5**	19.3	**62.3**	**53.8**	**66.7**
	G-CAM	**96.0**	30.9	0.1	**99.9**	5.3	0.3	**98.3**	1.9	3.7
Center-mass	LRP$_T$	49.9	**99.9**	15.4	**66.4**	**63.3**	**50.3**	**66.8**	**51.9**	28.8
	G-CAM	**81.0**	**66.3**	0.1	**67.3**	0.8	0.2	**72.7**	21.8	29.2

Fooling Success Rate (FSR) on 10,000 ImageNet validation set

Active Fooling : Swapping interpretations between classes

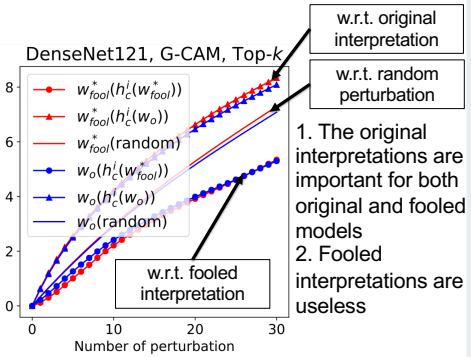


Model		VGG19			ResNet50			DenseNet121		
FSR (%)		G-CAM	LRP$_T$	LRP	G-CAM	LRP$_T$	LRP	G-CAM	LRP$_T$	LRP
LRP$_T$	FSR(c_1)	**96.5**	**94.5**	**97.0**	**90.5**	34.0	10.7	0.0	0.0	0.0
	FSR(c_2)	**96.5**	**95.0**	**96.0**	**75.0**	31.5	24.3	0.0	0.0	0.0
G-CAM	FSR(c_1)	1.0	0.0	1.0	**76.0**	0.0	0.0	4.0	0.0	0.0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

Accuracy and AOPC

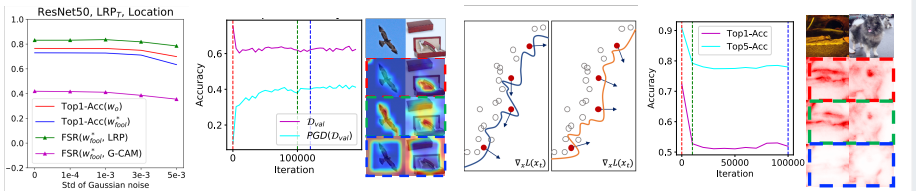
Model		VGG19		Resnet50		DenseNet121	
Accuracy (%)		Top1	Top5	Top1	Top5	Top1	Top5
Baseline (Pretrained)		72.4	90.9	76.1	92.9	74.4	92.0
Location	LRP$_T$	71.8	90.7	73.0	91.3	72.5	91.0
	G-CAM	71.5	90.4	74.2	91.8	73.7	91.6
Top-k	LRP$_T$	71.6	90.5	73.7	91.9	72.3	91.0
	G-CAM	72.1	90.6	74.7	92.0	73.1	91.2
Center 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

Prediction accuracy does not drop much



1. The original interpretations are important for both original and fooled models
2. Fooled interpretations are useless

Discussion



Robustness
Robust to Gaussian noise perturbation on weights

Adversarial training
Fooling still works for adversarially trained model

Intuition
Why fooling occurs

Limitation
SmoothGrad is difficult to fool