Fooling Neural Network Interpretations via Adversarial Model Manipulation

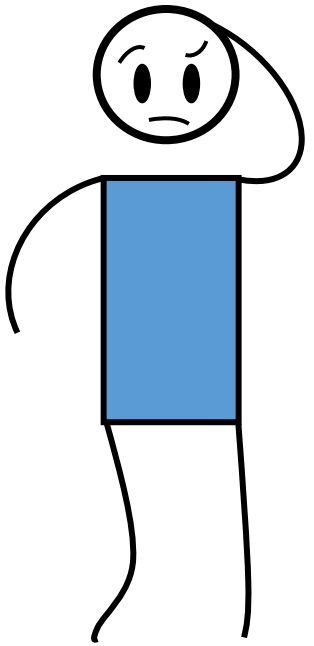
Juyeon Heo*[1] , Sunghwan Joo*[1,] and Taesup Moon[1,2]

Department of Electrical and Computer Engineering[1], Department of Artificial Intelligence[2] , Sungkyunkwan University, Korea
heojuyeon12@gmail.com, {shjoo840, tsmoon}@skku.edu

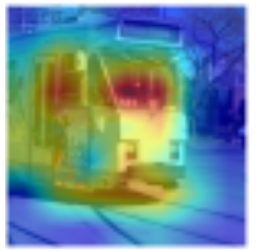
MOTIVATION

Both Reliability and Fairness are required in DNN

Complex Black Box model

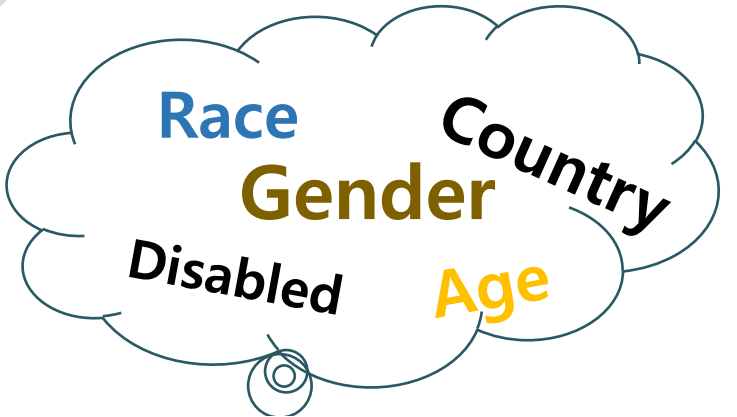


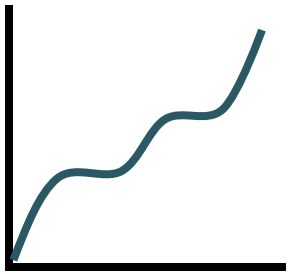
Reliability



or

Fairness





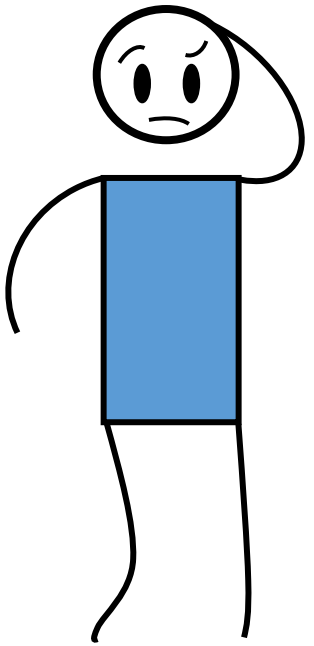
High Accuracy is
Not Enough for
validating the Model

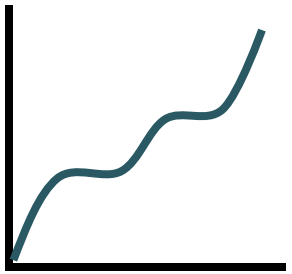
MOTIVATION

Both Reliability and Fairness are required in DNN

Complex Black Box model

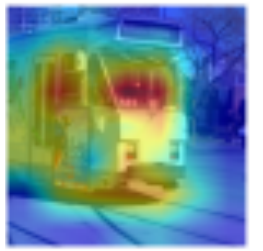
INTERPRETABLE METHOD





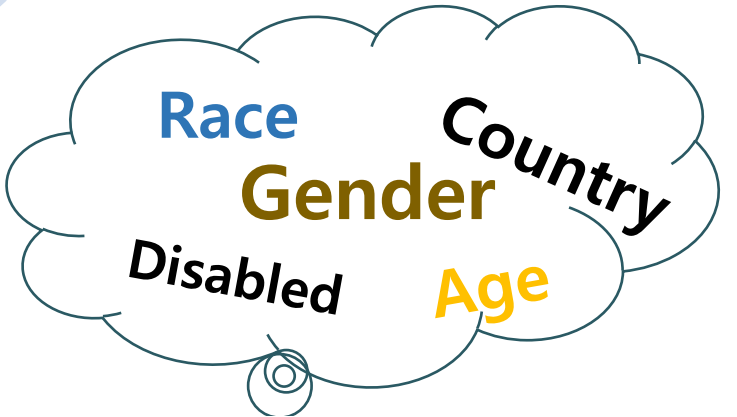
High Accuracy is Not Enough for validating the Model

Reliability



or

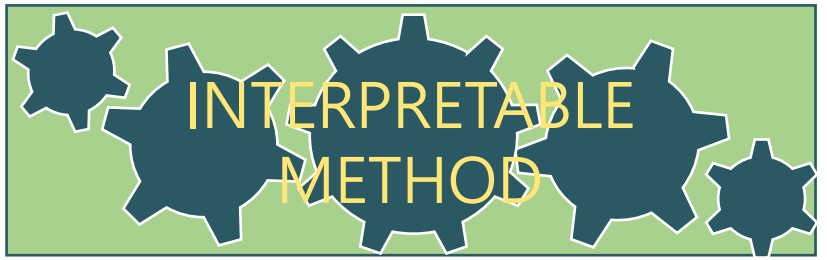
Fairness

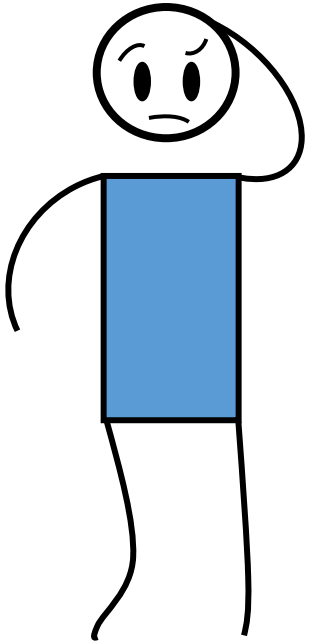


MOTIVATION

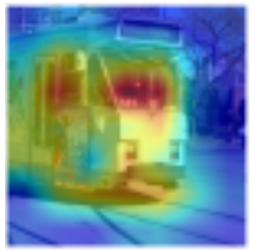
Both Reliability and Fairness are required in DNN

Complex Black Box model



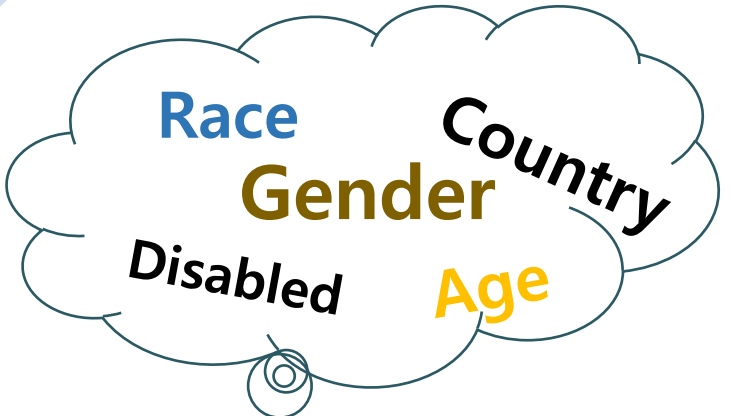


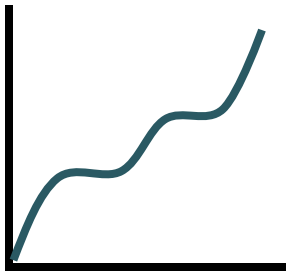
Reliability



or

Fairness



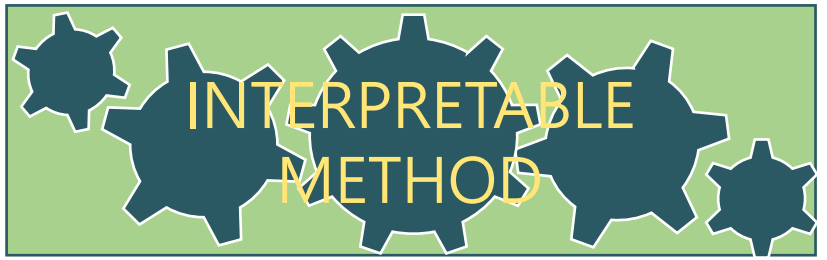


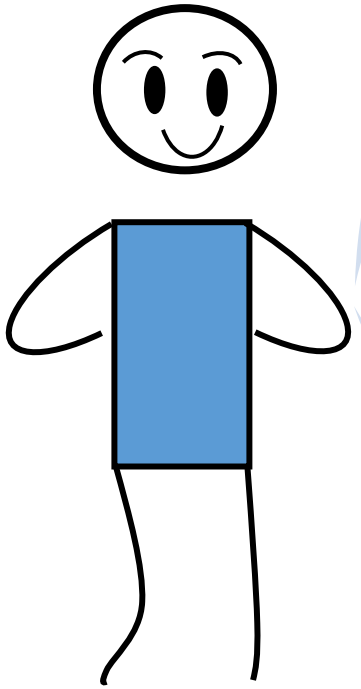
High Accuracy is
Not Enough for
validating the Model

MOTIVATION

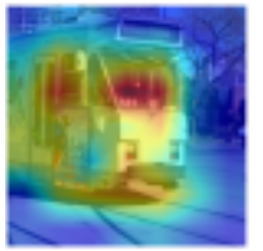
Both Reliability and Fairness are required in DNN

Complex Black Box model



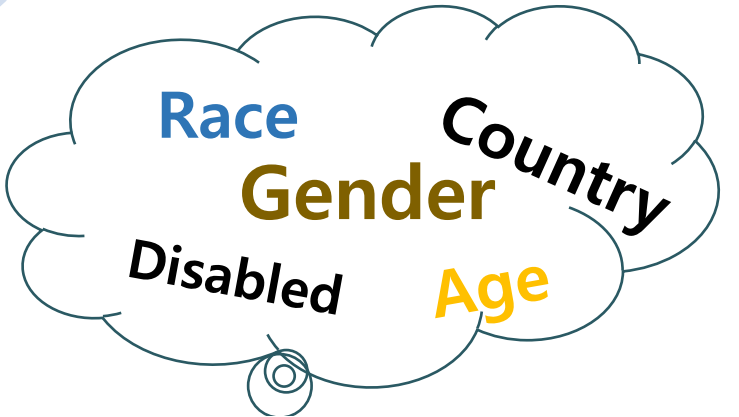


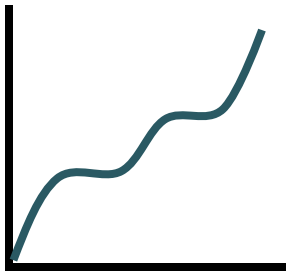
Reliability



or

Fairness



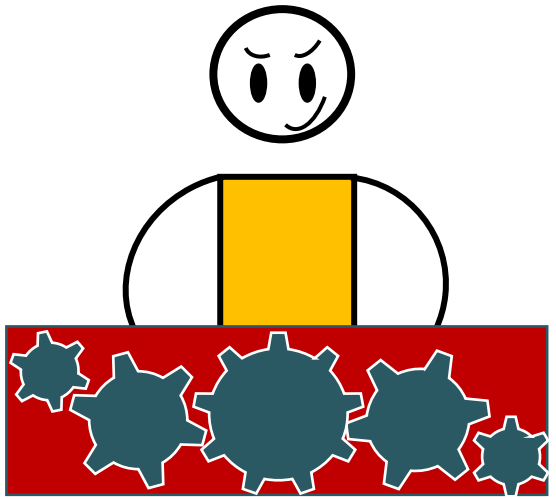


High Accuracy is
Not Enough for
validating the Model

MOTIVATION

Mean developer can manipulate the model to fool interpretations.

Mean Developer

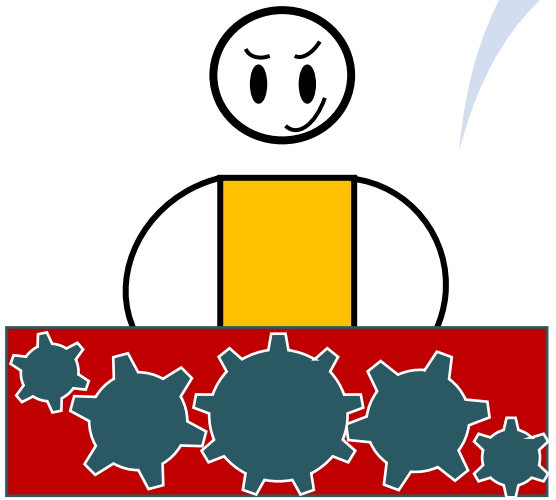


Fooling Interpretations via Model Manipulation!!

MOTIVATION

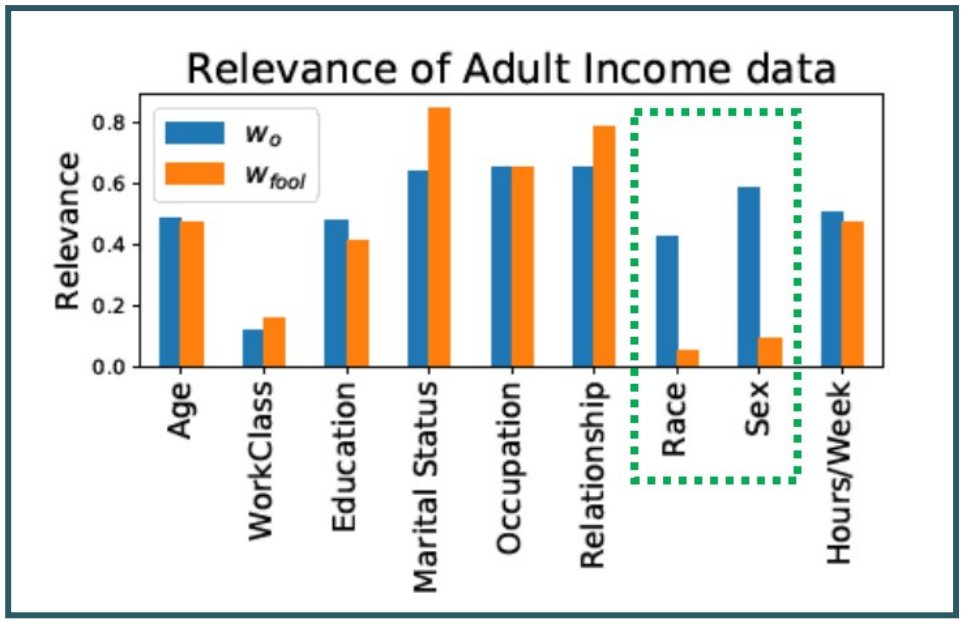
Mean developer can manipulate the model to fool interpretations.

Mean Developer



Fooling Interpretations via Model Manipulation!!

Fooled Interpretations!!

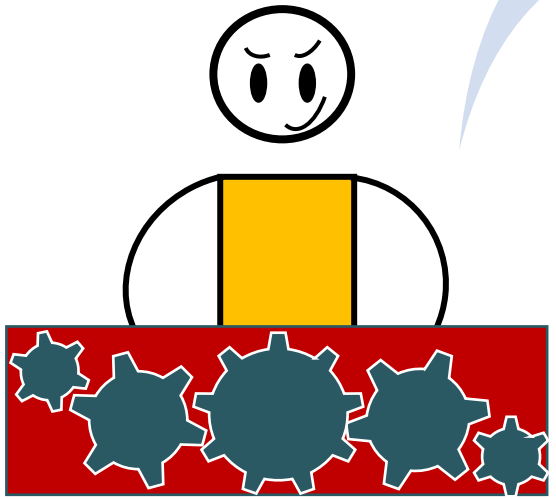


Try to hide the fact that model uses Race and Sex features

MOTIVATION

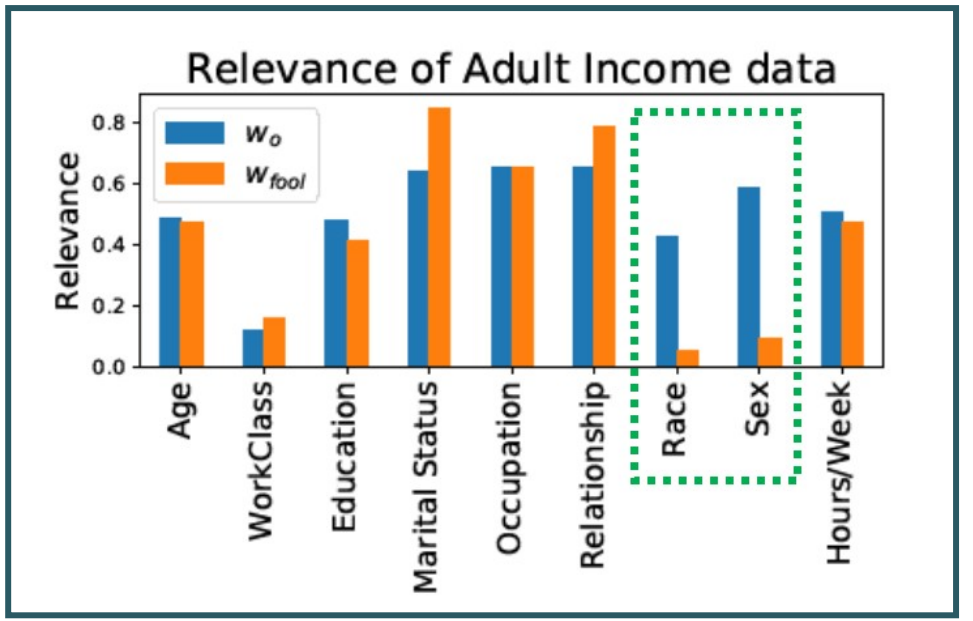
Mean developer can manipulate the model to fool interpretations.

Mean Developer

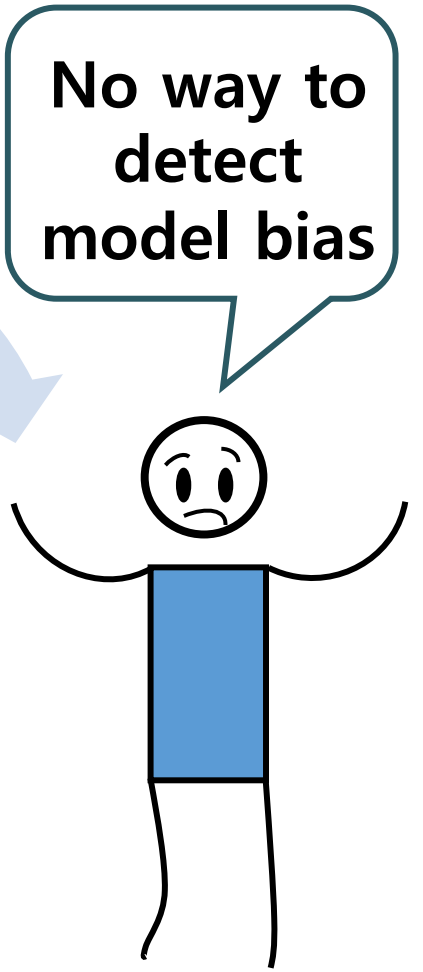


Fooling Interpretations via Model Manipulation!!

Fooled Interpretations!!



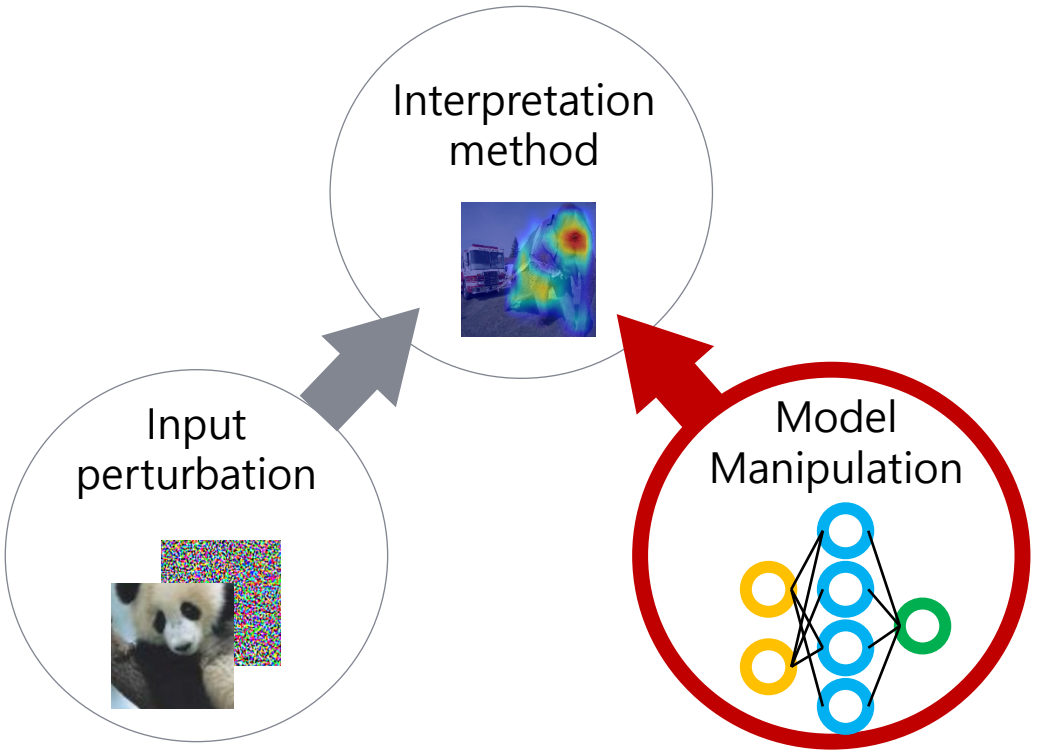
Try to hide the fact that model uses Race and Sex features



MOTIVATION

Our Fooling is generalizable without hurting accuracy

Methods for Fooling Interpretations



1. Our Fooling is **generalized to Entire validation set.**

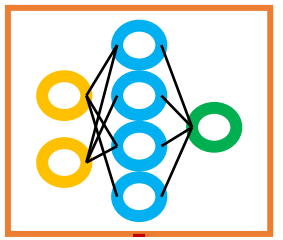
2. The accuracy drops around **only** 2% for Top-1 accuracy and 1% for Top-5 accuracy.

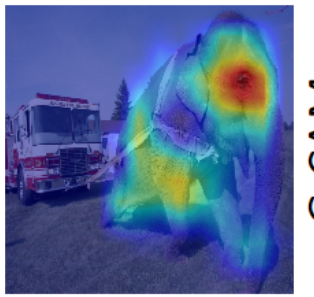
METHODS

Fine-tune trained model with Fooling Loss.

Ordinary Model

$(X, y, \mathcal{L}_{CE})$ Training



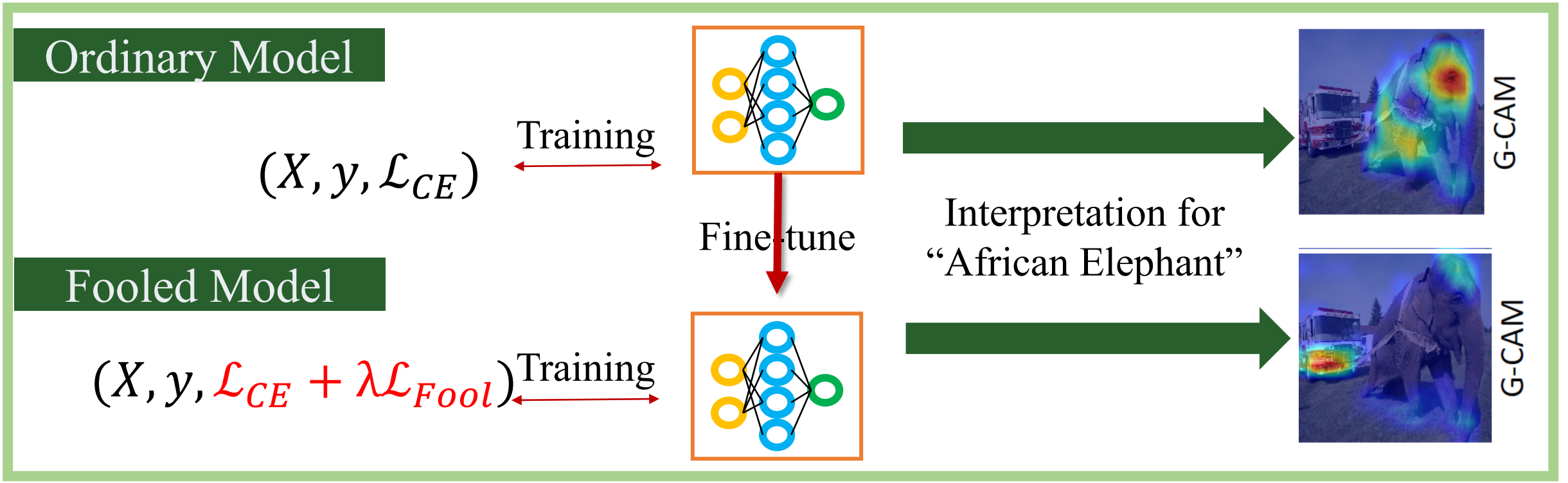


G-CAM

After training models with original cross entropy loss,
we **fine-tuned** the model with cross entropy loss and fooling loss.

METHODS

Fine-tune trained model with Fooling Loss.

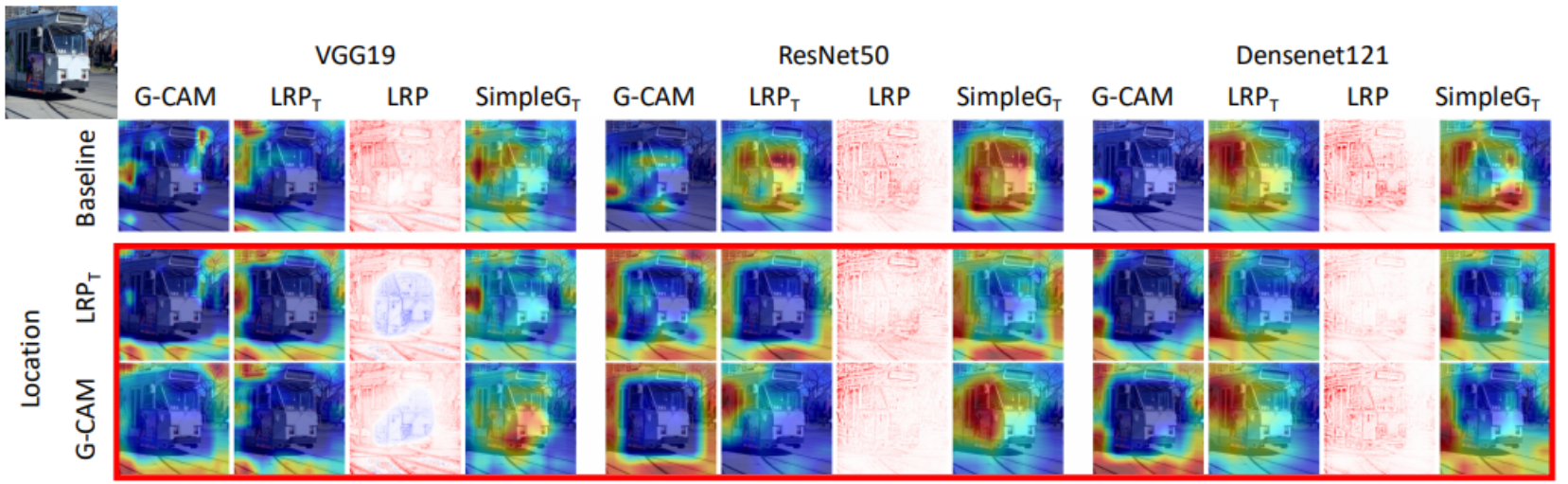


After training models with original cross entropy loss,
we **fine-tuned** the model with cross entropy loss and fooling loss.

RESULTS

Qualitative and Quantitative results of Fooling.

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**

- Fooling with LRP, interpretation with **LRP**

(1, 6) → (**2, 6**)

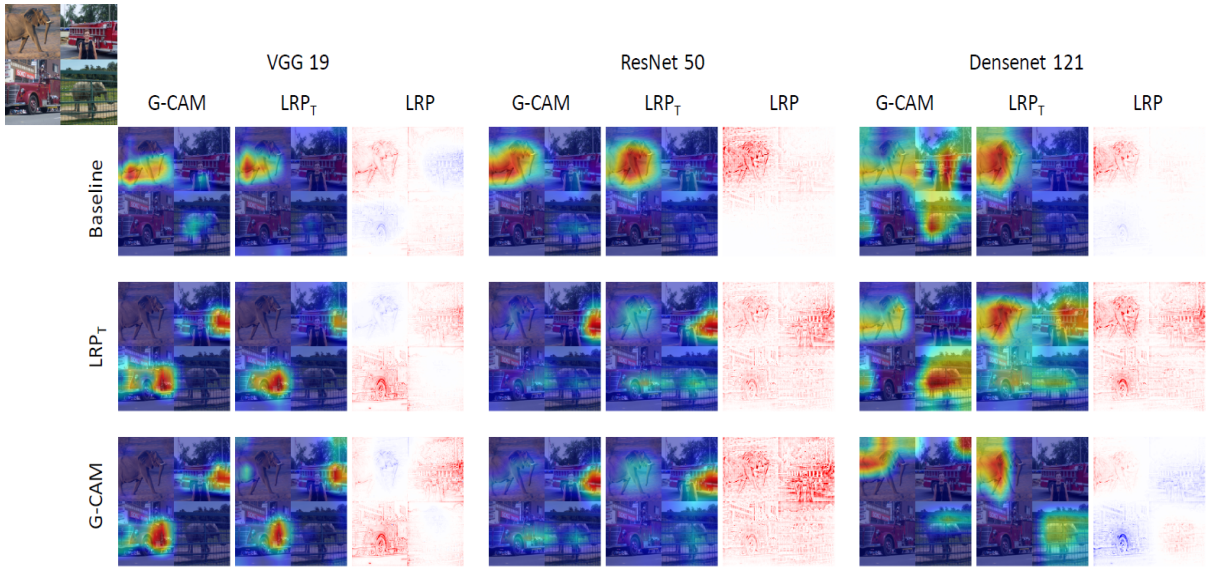
- Fooling with LRP, interpretation with **Others**

(1, 5:8) → (**2, 5:8**)

RESULTS

Qualitative and Quantitative results of Fooling.

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

- Fooling with Grad-CAM, interpretation with **Grad-CAM**

(1, 2) → (**2, 2**)

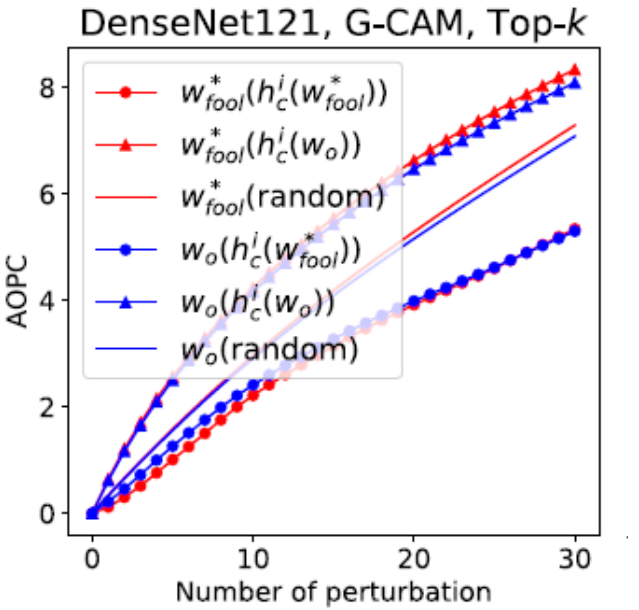
- Fooling with Grad-CAM, interpretation with **Others**

(1, 1:3) → (**3, 1:3**)

RESULTS

Accuracy after fooling and AOPC.

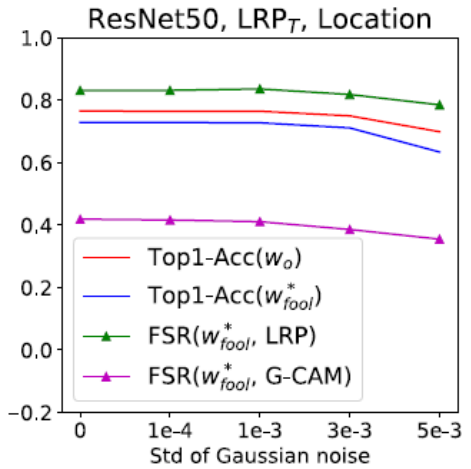
Model		VGG19		Resnet50		DenseNet121	
Accuracy (%)		Top1	Top5	Top1	Top5	Top1	Top5
Baseline (Pretrained)		72.4	90.9	76.1	92.9	74.4	92.0
Location	LRP_T	71.8	90.7	73.0	91.3	72.5	91.0
	G-CAM	71.5	90.4	74.2	91.8	73.7	91.6
Top-k	LRP_T	71.6	90.5	73.7	91.9	72.3	91.0
	G-CAM	72.1	90.6	74.7	92.0	73.1	91.2
Center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



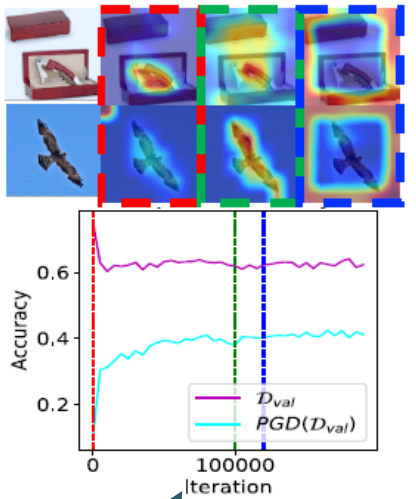
Accuracy drops **Around 2% in Top-1 and 1% in Top-5**
Both original model and fooled model focus **similar parts** where **original interpretations highlights** to make decisions.
Fooled interpretations are **not showing** what the model think.

RESULTS

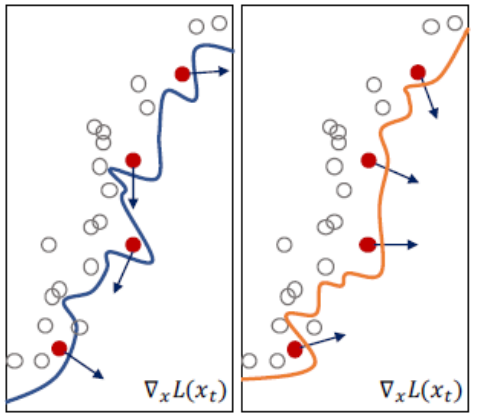
Why? And Extra experiments.



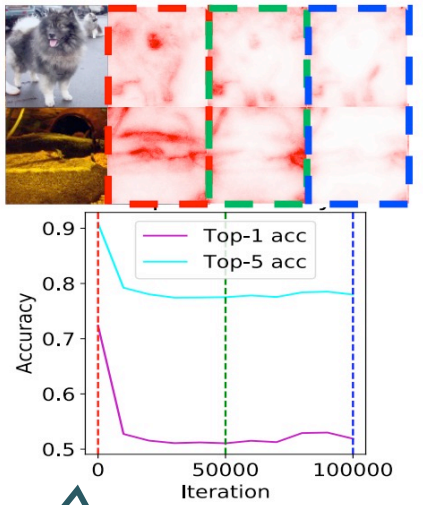
"Our foolings are robust to Gaussian noise addition on weights."



"Adversarially trained model can also be fooled by our foolings."



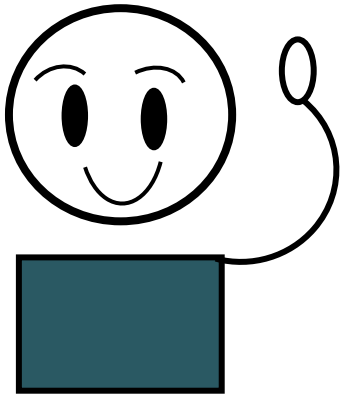
"Two decision boundaries can have almost the same accuracy, but different gradients, which lead to completely different interpretations."



(e) AMM by SG

"Hard to fool Smooth Grad while keeping accuracy. Hint to develop robust interpretation methods to our foolings."

Thank You



We hope this paper can open another research for more robust interpretation methods.