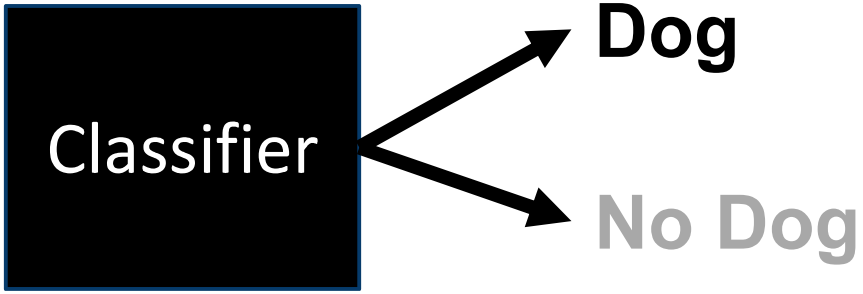
Adversarial Attacks - Explainability

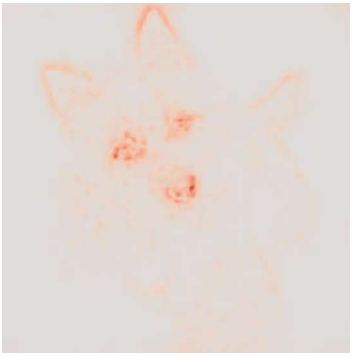
Dr. Fani Deligianni,
fani.deligianni@glasgow.ac.uk
Lecturer (Assistant Professor)
Lead of the Computing Technologies for Healthcare Theme
https://www.gla.ac.uk/schools/computing/staff/fanideligianni

Manipulated Explanations

Original Image

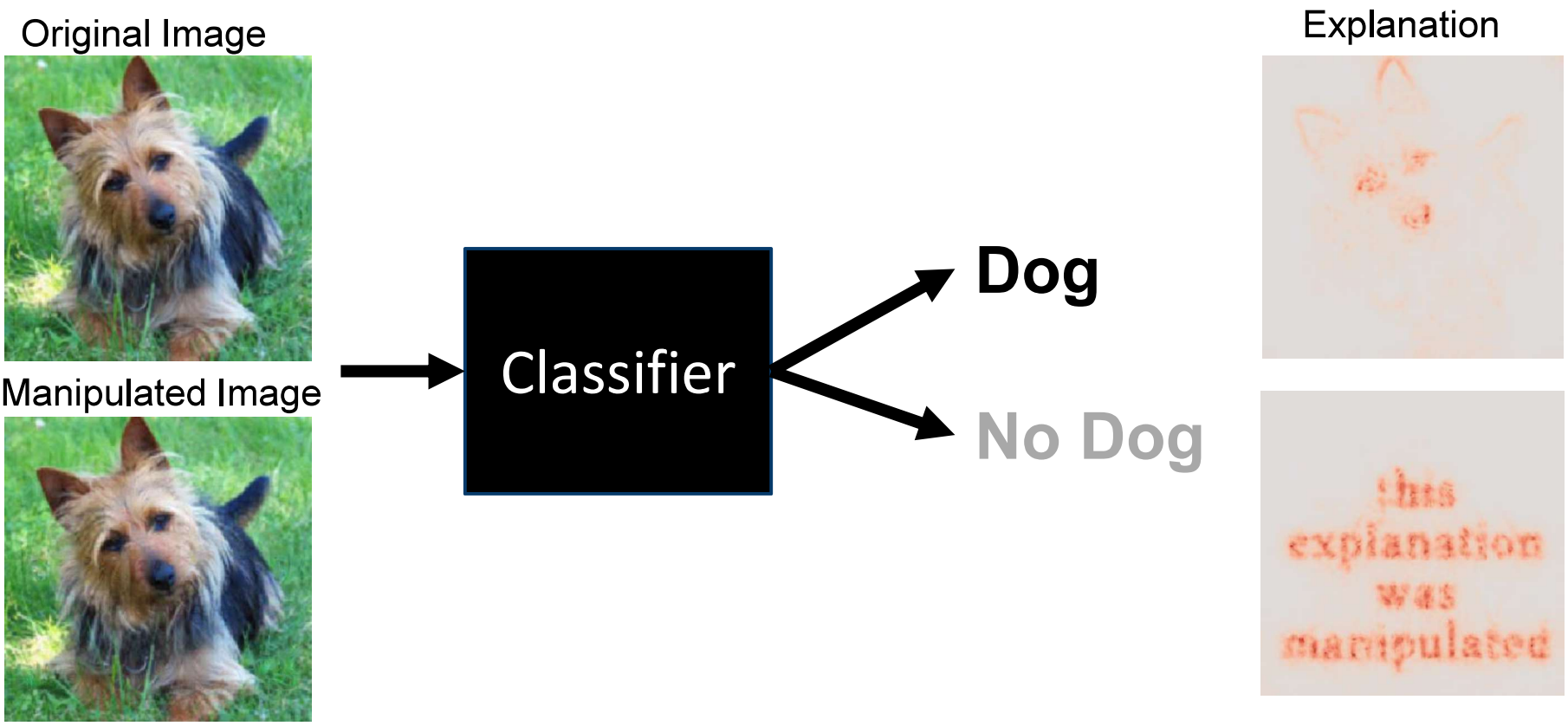


Explanation



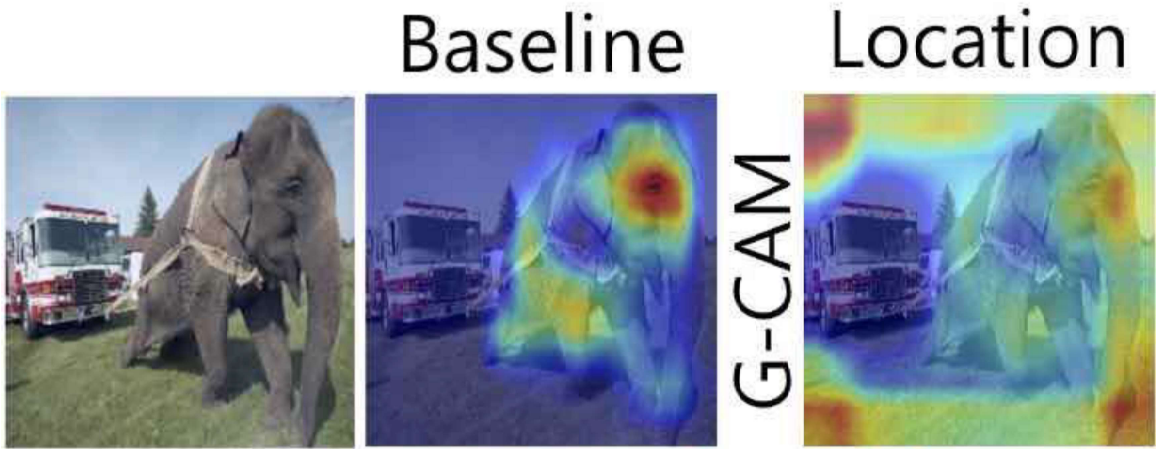
Dombrowski et al. ‘Explanations can be manipulated and geometry is to blame’, NeurIPS, 2019

Manipulated Explanations



Dombrowski et al. ‘Explanations can be manipulated and geometry is to blame’, NeurIPS, 2019

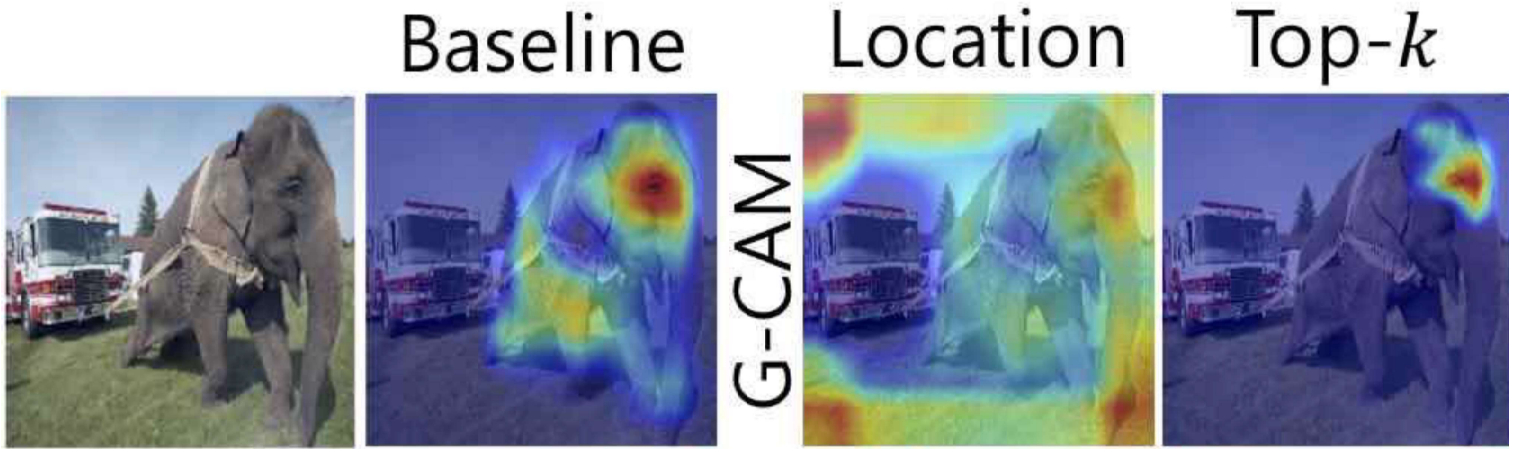
Model Manipulation



- Location fooling

Heo et al. ‘Fooling Neural Network Interpretations via Adversarial Model Manipulation’, NeurIPS, 2019

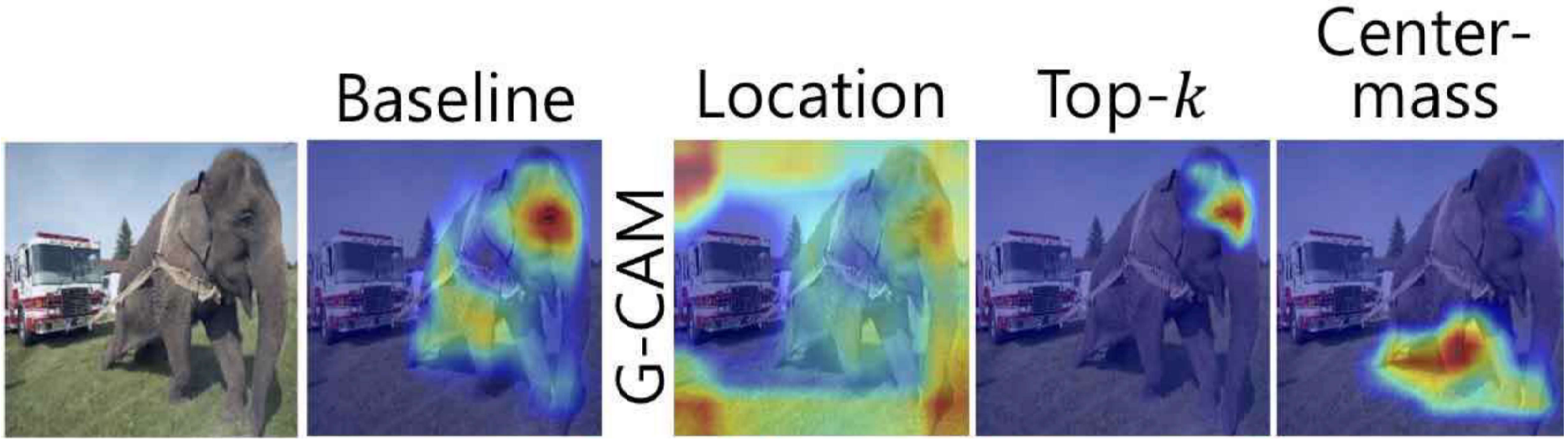
Model Manipulation



- Location fooling
- Top-k fooling

Heo et al. ‘Fooling Neural Network Interpretations via Adversarial Model Manipulation’, NeurIPS, 2019

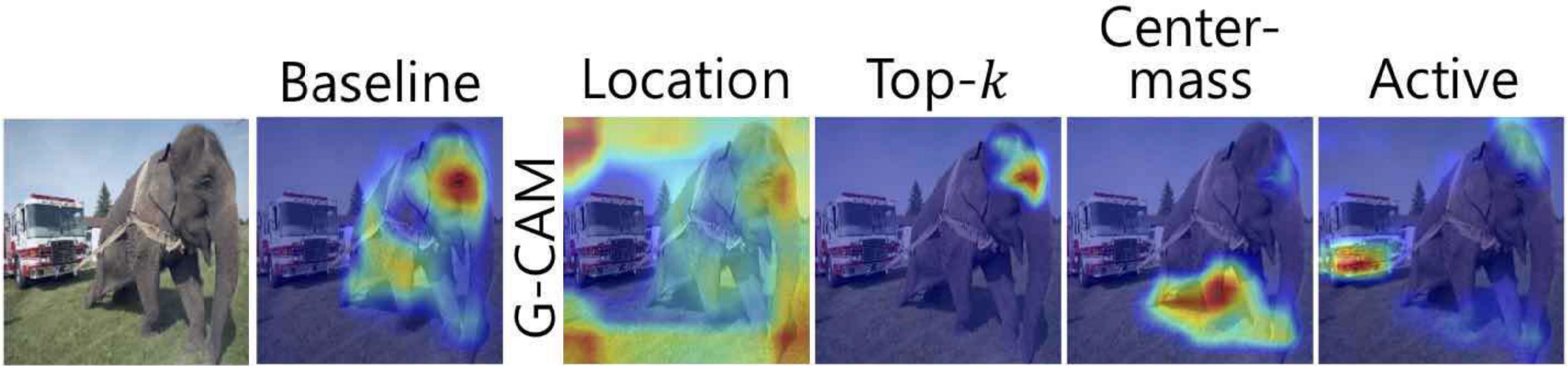
Model Manipulation



- Location fooling
- Top-k fooling
- Center-mass fooling

Heo et al. ‘Fooling Neural Network Interpretations via Adversarial Model Manipulation’, NeurIPS, 2019

Model Manipulation

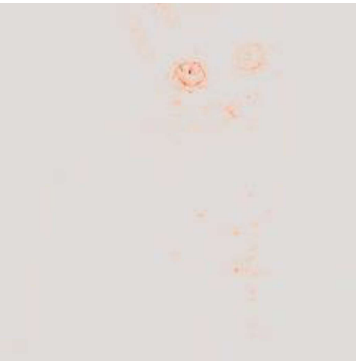


- Location fooling
- Top-k fooling
- Center-mass fooling
- Active fooling

Heo et al. ‘Fooling Neural Network Interpretations via Adversarial Model Manipulation’, NeurIPS, 2019

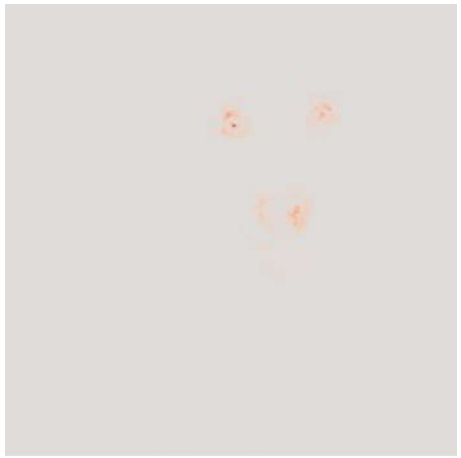
Data Perturbations

Target Image

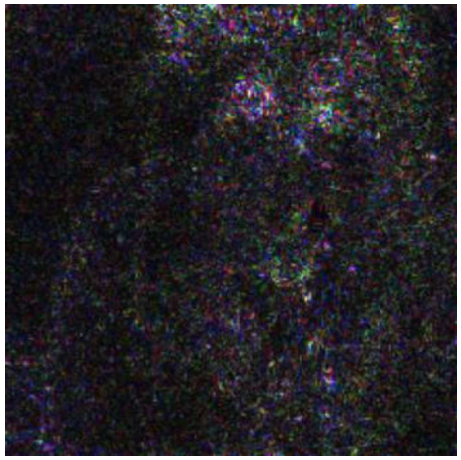


Target Explanations

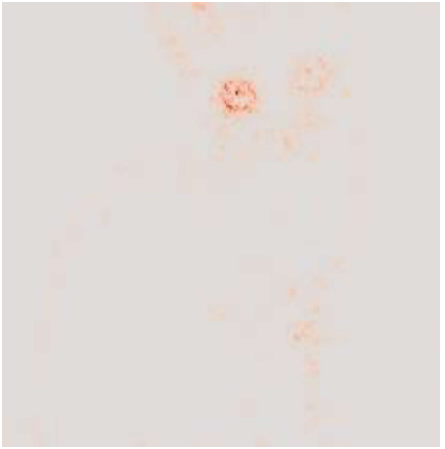
Original Image



Original Explanation



Perturbation

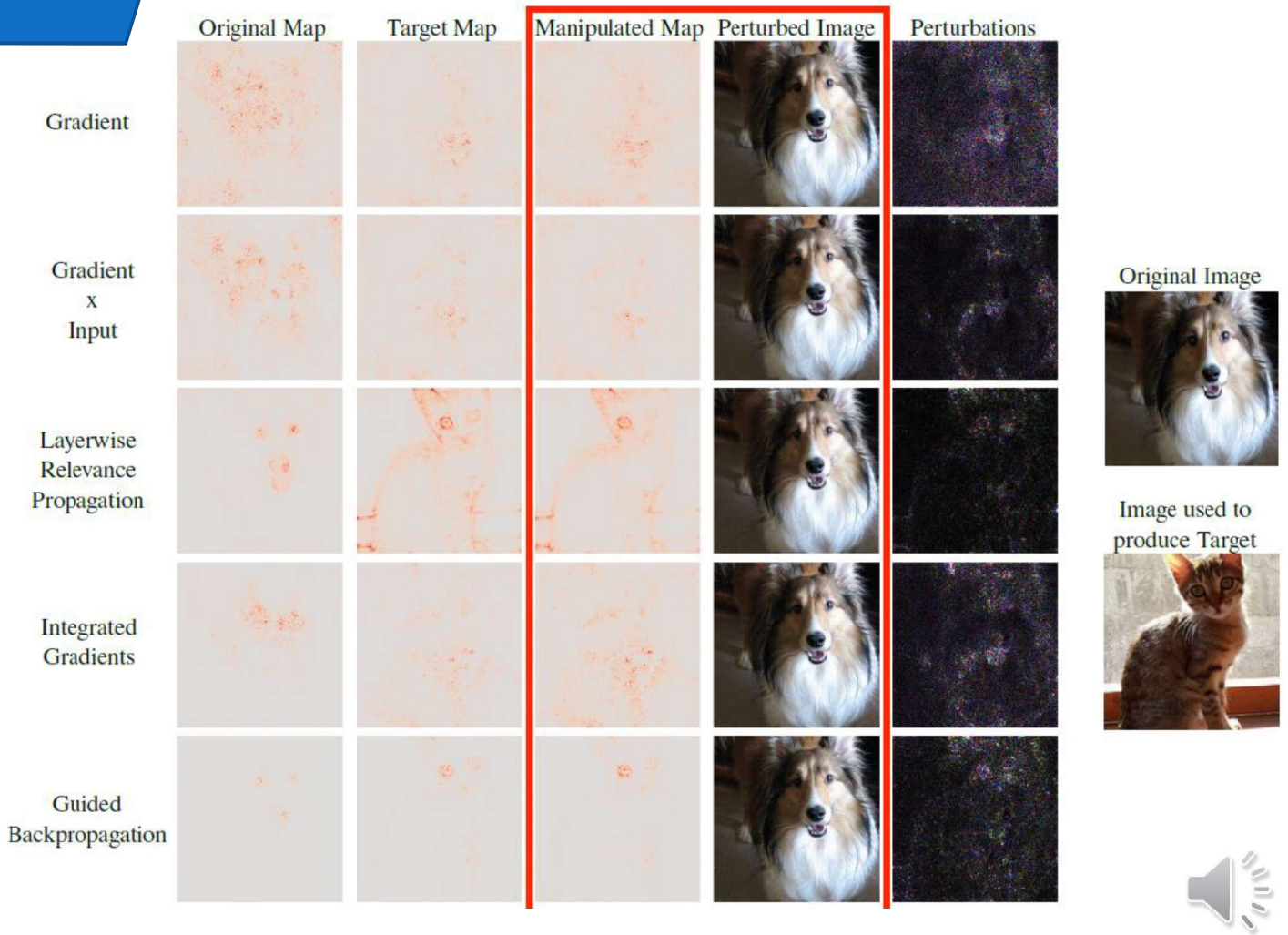


Manipulated Explanation

Dombrowski et al. ‘Explanations can be manipulated, and geometry is to blame’, NeurIPS, 2019

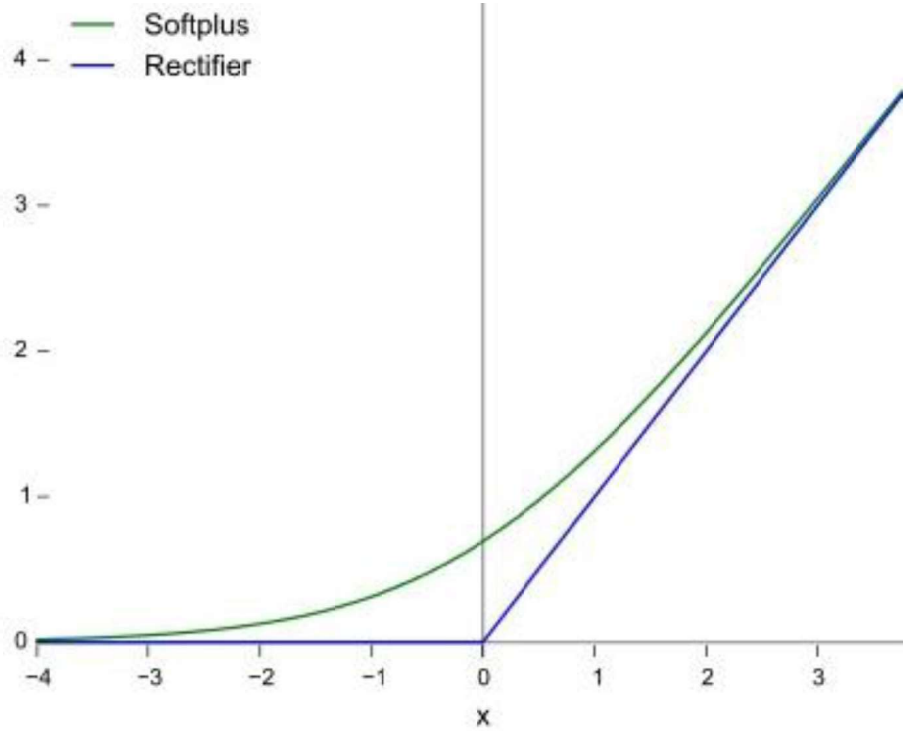
Example Perturbations

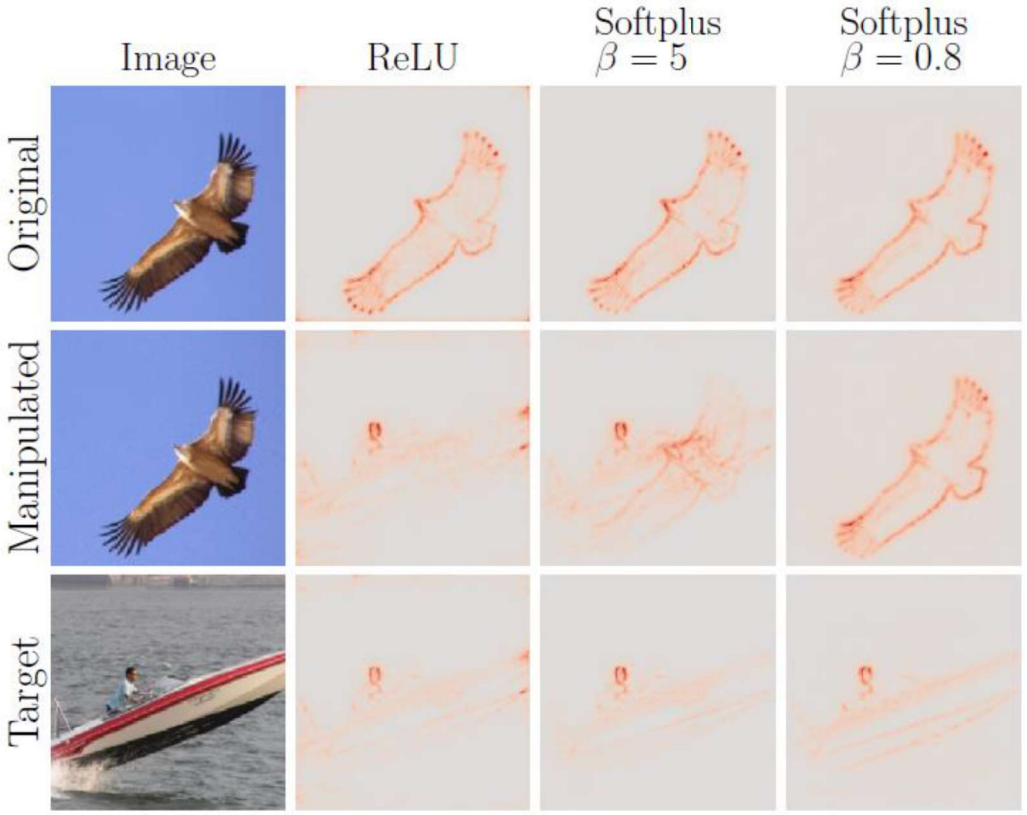
- Gradient-based methods are prone to adversarial attacks against explanations
- Where do we go from here?



Dombrowski et al. ‘Explanations can be manipulated, and geometry is to blame’, NeurIPS, 2019

Smoothing Explanations





Dombrowski et al. ‘Explanations can be manipulated, and geometry is to blame’, NeurIPS, 2019

Summary

- Gradient methods are affected from malicious attacks that can alter the explanation
- Explanations can be manipulated either directly by altering the model in a way that it introduces bias into explanations while it retains its accuracy
- Explanations can be manipulated by perturbing the input data in ways that are difficult to notice

References

- Dombrowski et al. ‘Explanations can be manipulated and geometry is to blame’, NeurIPS, 2019.
- Heo et al. ‘Fooling Neural Network Interpretations via Adversarial Model Manipulation’, NeurIPS, 2019.
- Subramanya et al. ‘Fooling Network Interpretation in Image Classification’, ICCV 2019.