When Causal Intervention meets Adversarial Examples and Image Masking

C.-H. Huck Yang*, Y.-C. Liu*, Pin-Yu Chen, Xiaoli Ma, and Yi-Chang James Tsai

**equal contribution*

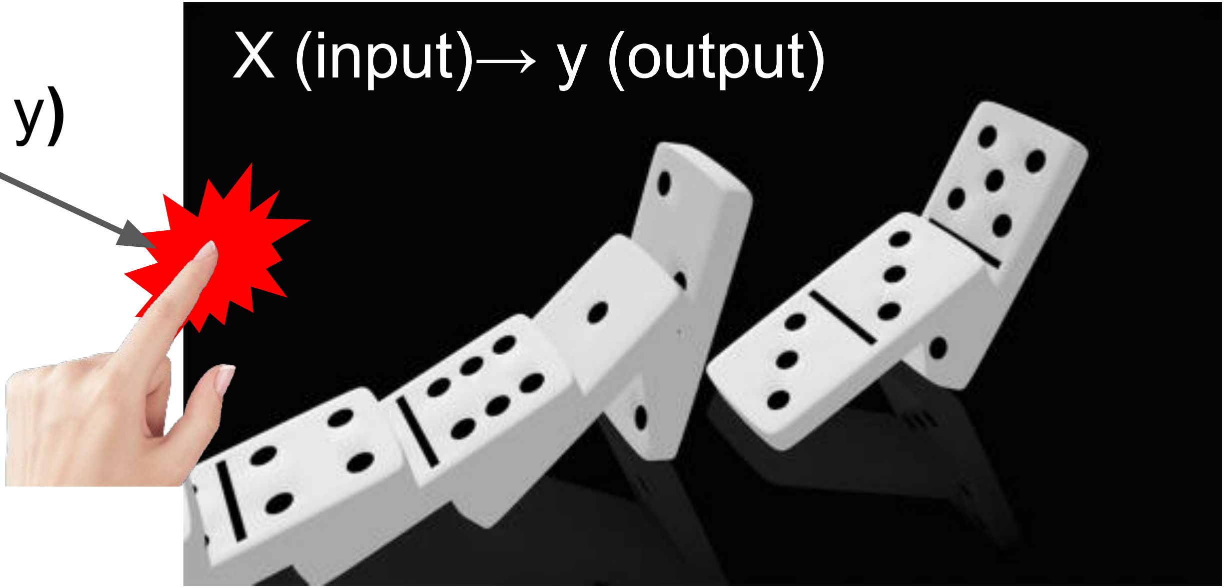
Georgia Institute of Technology

IBM Research, AI Foundation Group

What is the Causality?

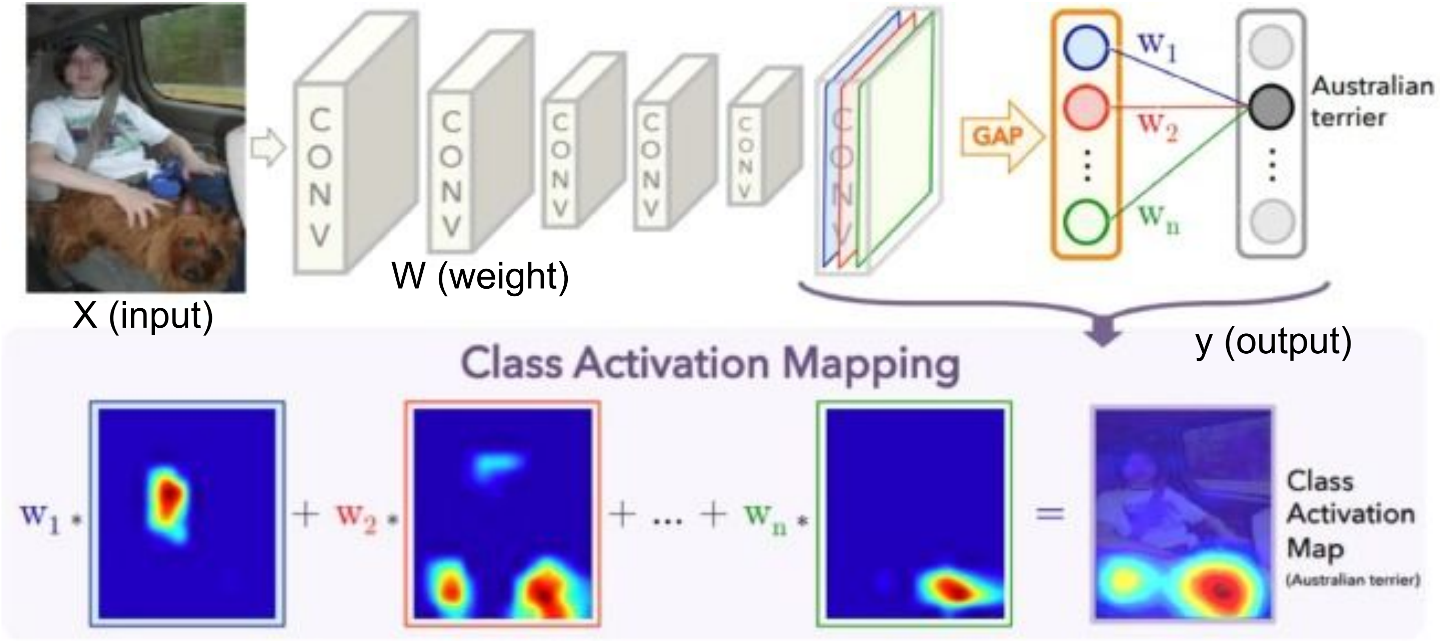
- Unidirectional

P(Causal Effect | X, y**)**



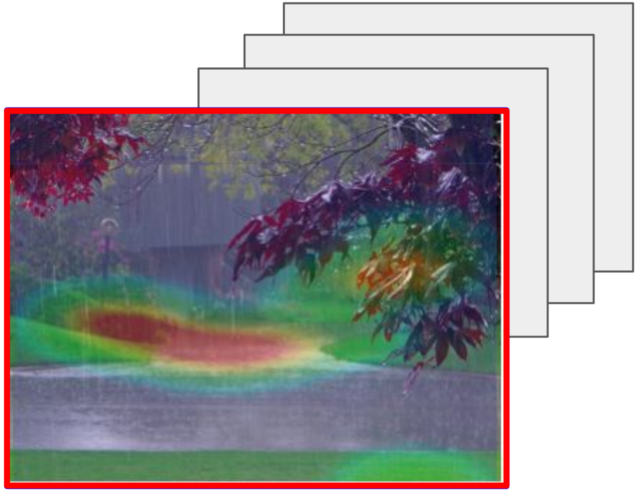
Source: Bechlivanidis C, Lagnado DA.

Related Work - Class-Activation Mapping (CAM)



Zhou et al. "Learning Deep Features for Discriminative Localization," CVPR, 2016

Correlation (CAM) v.s. Causality on DNNs



X (inputs)

Deep Neural Nets

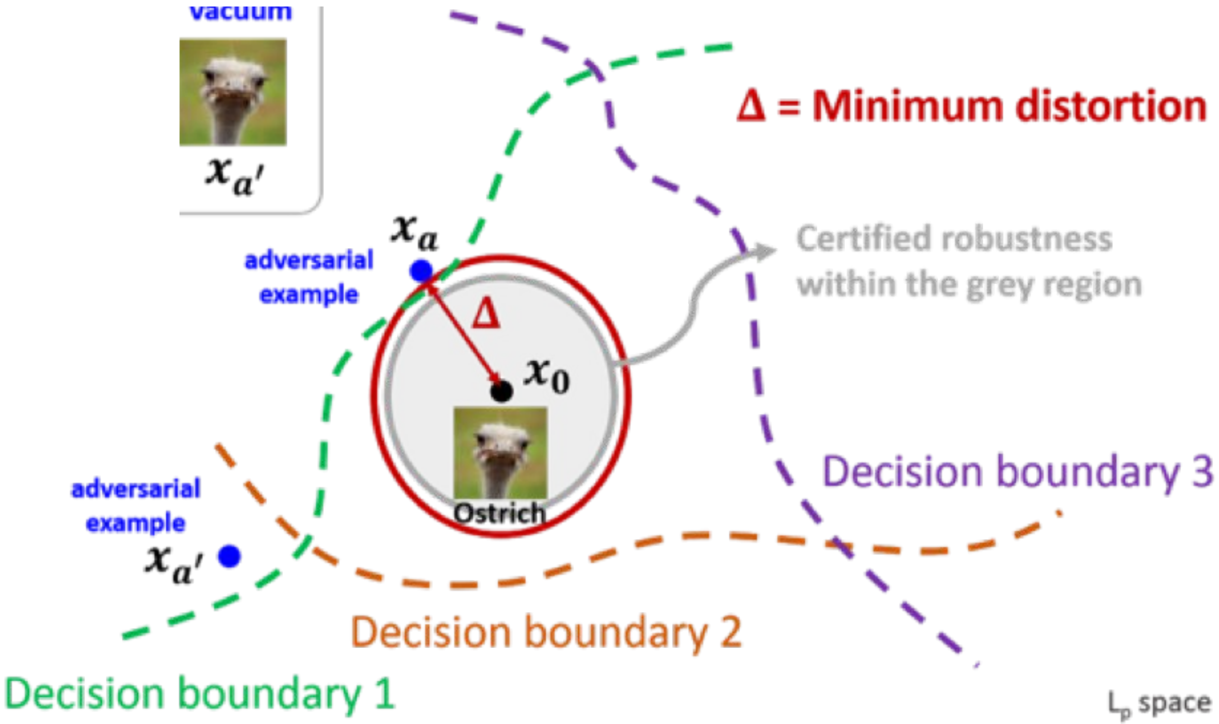
?

?

CAMs

how to find a visual causal features on output label?

New DNN Challenge: Adversarial Examples



Source: P.-Y. Chen, IBM Research

Proposed: Directed Causal Graph

Intervention : $do(x'_i)$

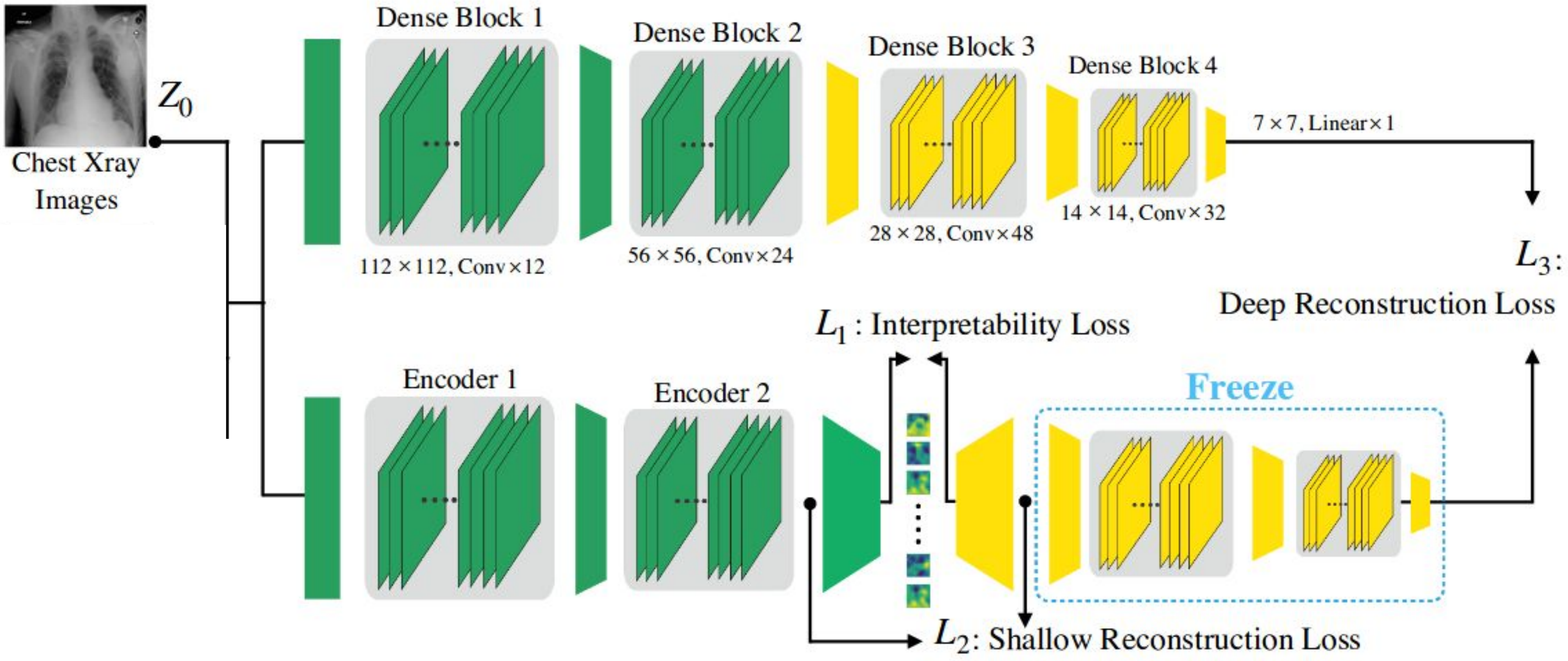
$$\boxed{Effect(x_i \rightarrow x_j, Z) = P(x_j | do(x'_i), Z_{X_i}) - P(x_j | Z_{X_i})}$$

Contributions

- Deep Autoencoding for calculating **causal effect (CE).**
- CE is a competitive **index** for understanding DNNs.
- We found that CE holds promises for detecting adversarial examples as it possesses distinct characteristics in the presence of adversarial perturbations.

Method: Deep Autoencoder

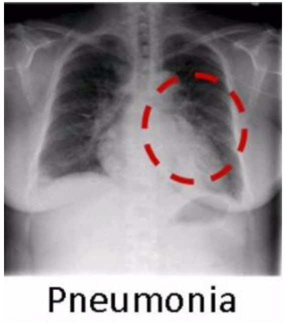
$$L(\theta; x_i) = \lambda_{shallow} \times L_{shallow}(\theta; x_i) + \lambda_{deep} \times L_{deep}(\theta; x_i) + \lambda_{interpretability} \times L_{interpretability}(\theta; x_i)$$



Harradon et al. “Causal Learning and Explanation of Deep Neural Networks via Autoencoded Activations” ,2018

Datasets

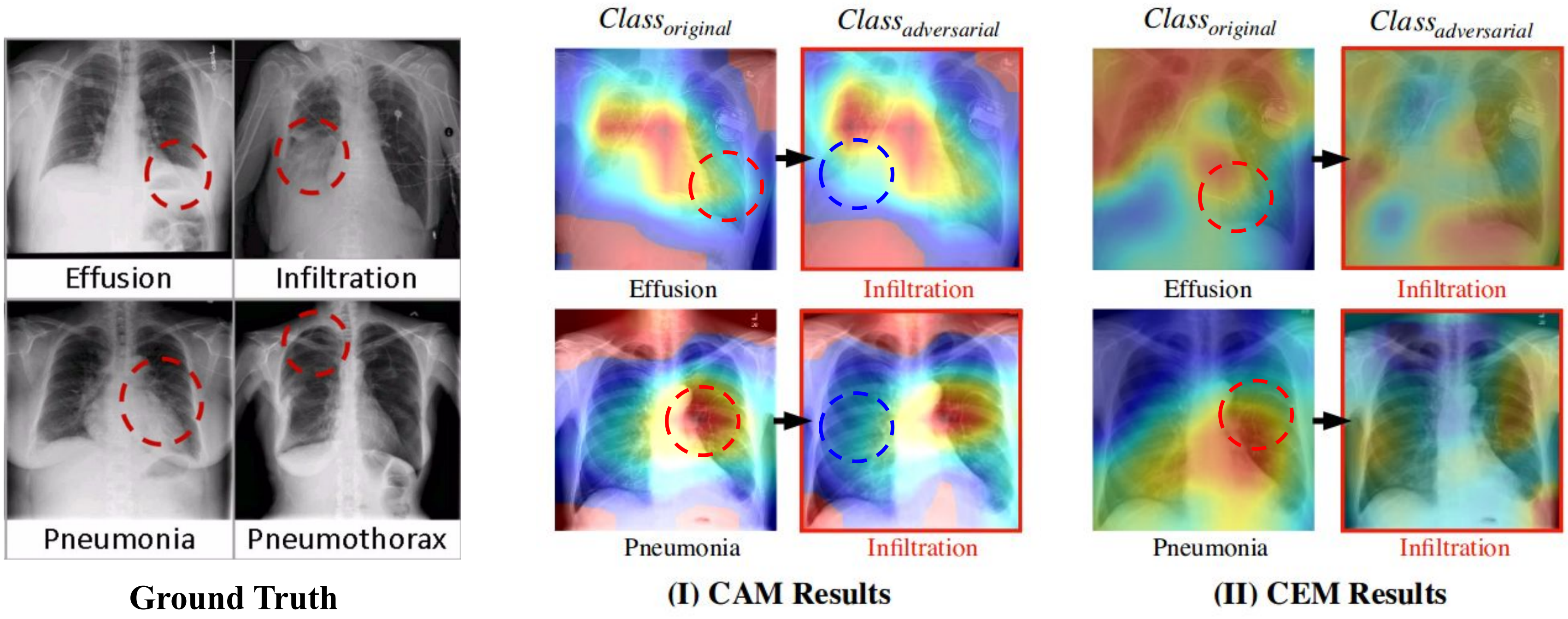
ChestX-ray14
- contains frontal-view chest X-ray
- 14 different thoracic diseases.



Fashion-MNIST
- 60k images

ImageNet
- 1.2 million images

Results - Causal Effect Mapping (CEM)



Estimate Causal Effect on Adversarial Signals

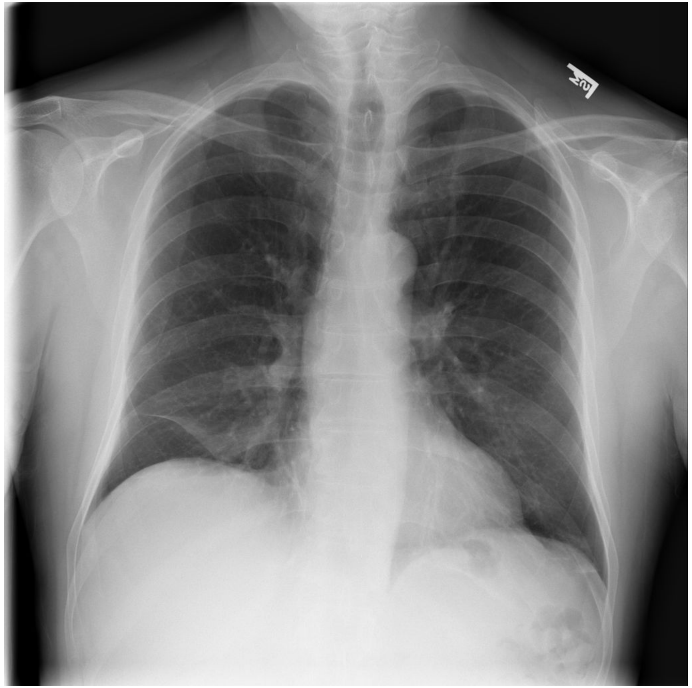


Table 1. Expected-CE on CheXNet

Level(L), Node(N)	Z_0	F_i = PWM
3,4	4.5076×10^{-3}	7.2356×10^{-7}
6,5	2.843×10^{-3}	1.2154×10^{-5}
6,10	3.1939×10^{-3}	9.0066×10^{-6}
8,5	3.1939×10^{-3}	1.1536×10^{-5}
10,7	1.3775×10^{-2}	-1.1506×10^{-5}

Estimate Causal Effect on Adversarial Signals

Attack Methods:

Fast Gradient Sign Method (FGSM)

Jacobian-Based Saliency Map (JBSM)

Basic Iterative Method (BIM)

Projected gradient descent (PGD)

Table 2. CheXNet (F_j on L = 10, N = 7)

F_j = Types of Adversarial Attack	Expected-CE
FSGM	**-5.6129**$\times 10^{-6}$
BIM	4.3435×10^{-5}
JBSM	7.7548×10^{-5}
PGD	-3.9605×10^{-6}

Conclusion

A Framework to estimate a causal effect on high dimensional visual data

Evaluate the this numerical causal effect on adversarial example

Saliency visualization as a interpertable method.

Future work:

Relation reasoning, time-series causal analysis, and video detection

Thank You!

Question & Answer

Code Released:
https://github.com/jjaacckkyy63/Causal-Intervention-AE-wAdvImg

yliu3233@gatech.com ; huckiyang@gatech.edu

Appendix: Expectation of Causal Effect

$X_j = x_j$ with all of the evidence Z could be computed as:

$$Effect(x_i \rightarrow x_j, Z) = P(x_j | do(x_i'), Z_{X_i}) - P(x_j | Z_{X_i}) \tag{1}$$

The excepted casual effect from **Eqn. 6** in [18] has been defined as:

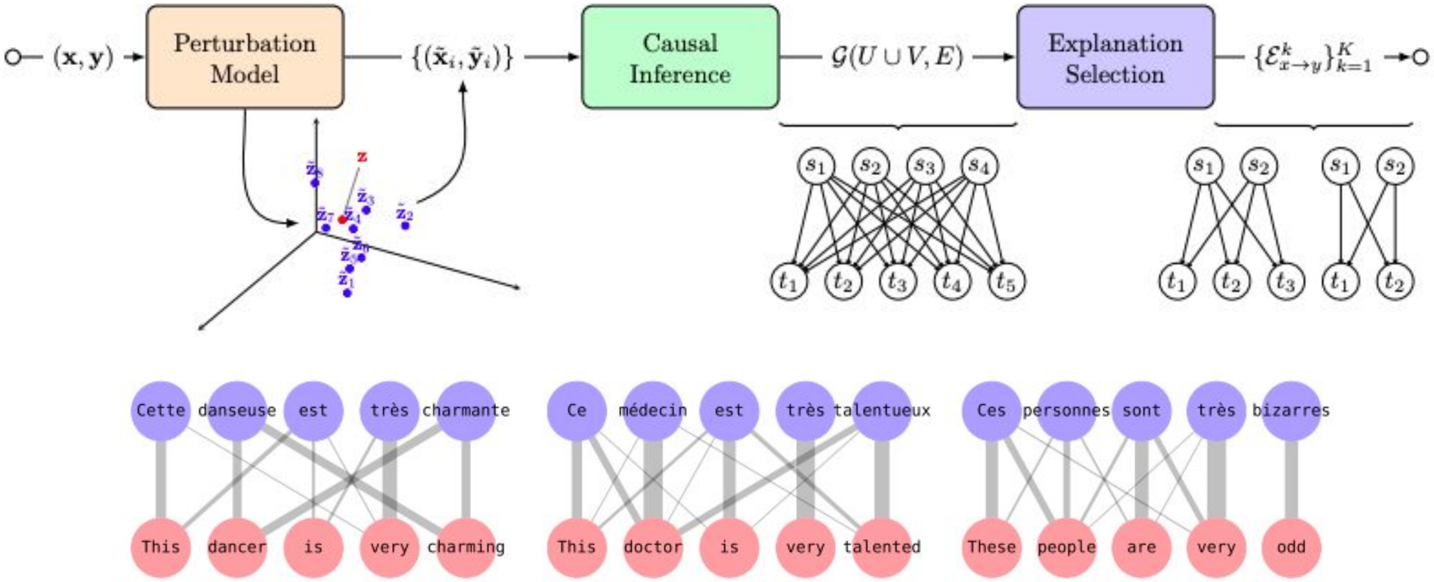
$$E_{X_i}[Effect(x_i \rightarrow x_j, Z)] = \sum_{x_i \in X_i} P(X_i = x_i | Z) \times (1) \tag{2}$$

$$P(x_i | pa_i') = \begin{cases} P(x_i | pa_i) & if\ F_1 = idle, \\ 0 & if\ F_i = do(x_i')\ and\ x_i \neq x_i', \\ 1 & if\ F_i = do(x_i')\ and\ x_i = x_i'. \end{cases} \tag{3}$$

$$L(\theta; x_i) = \lambda_{shallow} \times L_{shallow}(\theta; x_i) + \lambda_{deep} \times L_{deep}(\theta; x_i) + \lambda_{interpretability} \times L_{interpretability}(\theta; x_i) \tag{4}$$

Appendix - NLP

A causal framework for explaining the predictions of black-box sequence-to-sequence models, David Alvarez-Melis, Tommi S. Jaakkola, ACL, 2017



Explanations for biased translations of similar gender-neutral English sentences into French.

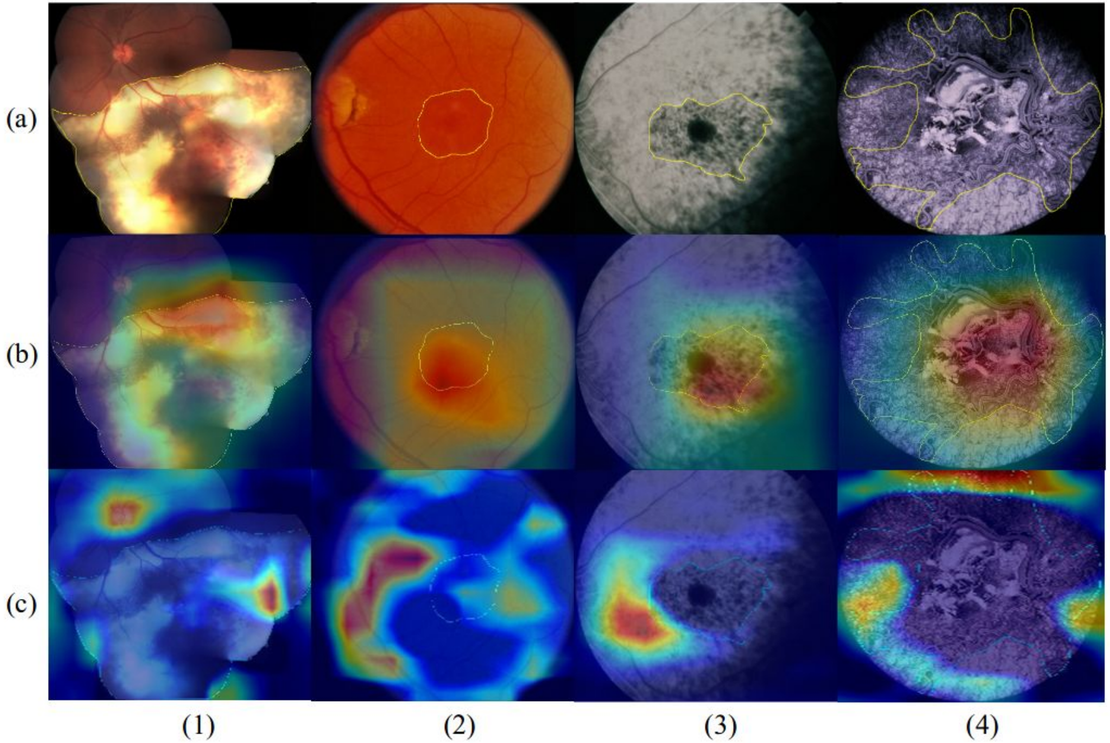
Simularity

- Biscuit v.s. Dog (Chihuahua)

Source: Mariya Yao 2018

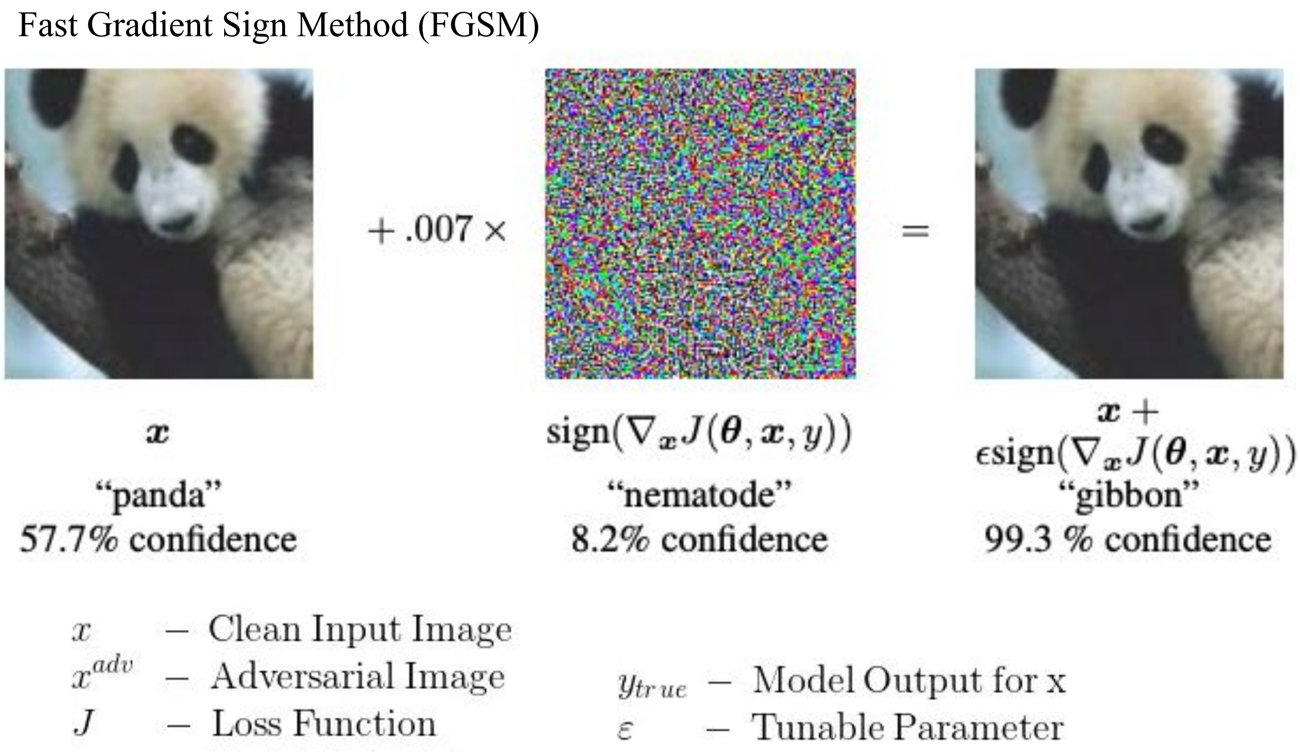
Medical Image

- Retinal Images



Source: [2] Yang et al. 2018 Biology Workshop ICML

Medical Image



Source: [3] Goodfellow et al. arxiv 2015