When Causal Intervention meets Adversarial Examples and Image Masking

*equal contribution
Georgia Institute of Technology
IBM Research, AI Foundation Group
What is the Causality?
- Unidirectional

\[ P(\text{Causal Effect} \mid X, y) \]

Source: Bechlivanidis C, Lagnado DA.
Related Work - Class-Activation Mapping (CAM)

Correlation (CAM) v.s. Causality on DNNs

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: \( \text{do}(x'_i) \)

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

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

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)

Ground Truth

(I) CAM Results

(II) CEM Results
Estimate Causal Effect on Adversarial Signals

<table>
<thead>
<tr>
<th>Level(L), Node(N)</th>
<th>$Z_0$</th>
<th>$F_i = \text{PWM}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3,4</td>
<td>4.5076 × 10^{-3}</td>
<td>7.2356 × 10^{-7}</td>
</tr>
<tr>
<td>6,5</td>
<td>2.843 × 10^{-3}</td>
<td>1.2154 × 10^{-5}</td>
</tr>
<tr>
<td>6,10</td>
<td>3.1939 × 10^{-3}</td>
<td>9.0066 × 10^{-6}</td>
</tr>
<tr>
<td>8,5</td>
<td>3.1939 × 10^{-3}</td>
<td>1.1536 × 10^{-5}</td>
</tr>
<tr>
<td>10,7</td>
<td>1.3775 × 10^{-2}</td>
<td>-1.1506 × 10^{-5}</td>
</tr>
</tbody>
</table>
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 $L = 10$, $N = 7$)

<table>
<thead>
<tr>
<th>$F_j$ = Types of Adversarial Attack</th>
<th>Expected-CE</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSGM</td>
<td>$-5.6129 \times 10^{-6}$</td>
</tr>
<tr>
<td>BIM</td>
<td>$4.3435 \times 10^{-5}$</td>
</tr>
<tr>
<td>JBSM</td>
<td>$7.7548 \times 10^{-5}$</td>
</tr>
<tr>
<td>PGD</td>
<td>$-3.9605 \times 10^{-6}$</td>
</tr>
</tbody>
</table>
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 interpretable 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:

\[
\text{Effect}(x_i \rightarrow x_j, Z) = P(x_j | do(x_i'), Z) - P(x_j | Z_{X_i})
\]  \hspace{1cm} (1)

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

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

\[
P(x_i | pa_i') = \begin{cases} 
P(x_i | pa_i) & \text{if } F_1 = \text{idle}, \\
0 & \text{if } F_i = do(x_i') \text{ and } x_i \neq x_i', \\
1 & \text{if } F_i = do(x_i') \text{ and } x_i = x_i'.
\end{cases}
\]  \hspace{1cm} (3)

\[
L(\theta; x_i) = \lambda_{\text{shallow}} \times L_{\text{shallow}}(\theta; x_i) + \lambda_{\text{deep}} \times L_{\text{deep}}(\theta; x_i) + \lambda_{\text{interpretability}} \times L_{\text{interpretability}}(\theta; x_i)
\]  \hspace{1cm} (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

Fast Gradient Sign Method (FGSM)

\[ x = \text{"panda" \ 57.7\% confidence} \]

\[ x^{adv} = \text{sign}(\nabla_x J(\theta, x, y)) = \text{"nematode" \ 8.2\% confidence} \]

\[ x + \epsilon \text{sign}(\nabla_x J(\theta, x, y)) = x^{adv} = \text{"gibbon" \ 99.3\% confidence} \]

- $x$ — Clean Input Image
- $x^{adv}$ — Adversarial Image
- $J$ — Loss Function
- $y_{true}$ — Model Output for $x$
- $\epsilon$ — Tunable Parameter