Generative causal explanations of black-box classifiers

Authors: Matthew O'Shaughnessy, Gregory Canal, Marissa Connor, Mark Davenport, and Christopher Rozell

Avinash Kori
Supervisor: Francesca Toni and Ben Glocker

Imperial College London

March 20, 2022

Structure

- Brief introduction to Vatriational AutoEncoders (VAE)[1] and Causal Inference[2].

[1] Diederik P Kingma and Max Welling. "An introduction to variational autoencoders". In: *arXiv preprint arXiv:1906.02691* (2019).

[2] Judea Pearl. "The seven tools of causal inference, with reflections on machine learning". In: *Communications of the ACM* 62.3 (2019), pp. 54–60.

Structure

- Brief introduction to Vatriational AutoEncoders (VAE)[1] and Causal Inference[2].
- Relevance to STAI.

[1] Diederik P Kingma and Max Welling. "An introduction to variational autoencoders". In: *arXiv preprint arXiv:1906.02691* (2019).

[2] Judea Pearl. "The seven tools of causal inference, with reflections on machine learning". In: *Communications of the ACM* 62.3 (2019), pp. 54–60.

Structure

- Brief introduction to Vatriational AutoEncoders (VAE)[1] and Causal Inference[2].
- Relevance to STAI.
- Main contributions in the paper.

[1] Diederik P Kingma and Max Welling. "An introduction to variational autoencoders". In: *arXiv preprint arXiv:1906.02691* (2019).

[2] Judea Pearl. "The seven tools of causal inference, with reflections on machine learning". In: *Communications of the ACM* 62.3 (2019), pp. 54–60.

Structure

- Brief introduction to Vatriational AutoEncoders (VAE)[1] and Causal Inference[2].
- Relevance to STAI.
- Main contributions in the paper.
- Methodology

[1] Diederik P Kingma and Max Welling. "An introduction to variational autoencoders". In: *arXiv preprint arXiv:1906.02691* (2019).

[2] Judea Pearl. "The seven tools of causal inference, with reflections on machine learning". In: *Communications of the ACM* 62.3 (2019), pp. 54–60.

Structure

- Brief introduction to Vatriational AutoEncoders (VAE)[1] and Causal Inference[2].
- Relevance to STAI.
- Main contributions in the paper.
- Methodology
- Results

[1] Diederik P Kingma and Max Welling. "An introduction to variational autoencoders". In: *arXiv preprint arXiv:1906.02691* (2019).

[2] Judea Pearl. "The seven tools of causal inference, with reflections on machine learning". In: *Communications of the ACM* 62.3 (2019), pp. 54–60.

Structure

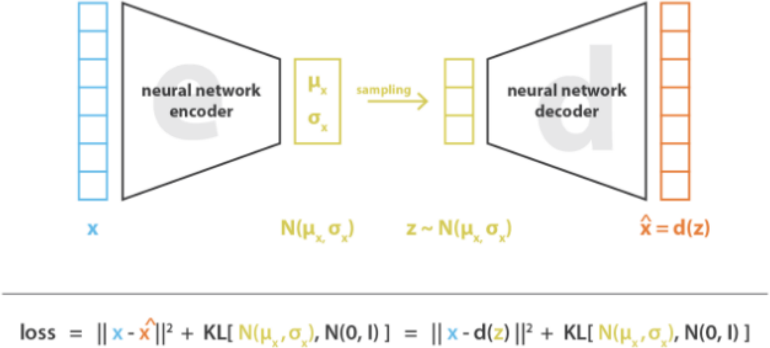
- Brief introduction to Vatriational AutoEncoders (VAE)[1] and Causal Inference[2].
- Relevance to STAI.
- Main contributions in the paper.
- Methodology
- Results
- Limitations and Future directions

[1] Diederik P Kingma and Max Welling. "An introduction to variational autoencoders". In: *arXiv preprint arXiv:1906.02691* (2019).

[2] Judea Pearl. "The seven tools of causal inference, with reflections on machine learning". In: *Communications of the ACM* 62.3 (2019), pp. 54–60.

Vatriational AutoEncoders

VAEs are generative methods to approximate the data distribution with an explicit likelihood formulation.



Vatriational AutoEncoders

Key advantages and uses for VAE

- VAE approximates data distribution.

[3] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. 2009.

Vatriational AutoEncoders

Key advantages and uses for VAE

- VAE approximates data distribution.
- VAE are easy to train when compared to any model $\in$ GAN family.

[3] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. 2009.

Vatriational AutoEncoders

Key advantages and uses for VAE

- VAE approximates data distribution.
- VAE are easy to train when compared to any model $\in$ GAN family.
- VAE framework helps us to control latent space, in terms of disentanglement.

[3] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. 2009.

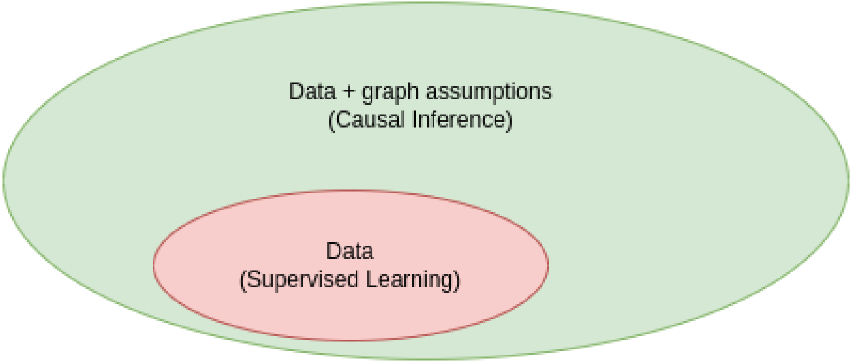
Vatriational AutoEncoders

Key advantages and uses for VAE

- VAE approximates data distribution.
- VAE are easy to train when compared to any model $\in$ GAN family.
- VAE framework helps us to control latent space, in terms of disentanglement.
- As VAEs assume Gaussian priors, this results in elegant bound estimation (further reading: ELBO[3]).

[3] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. 2009.

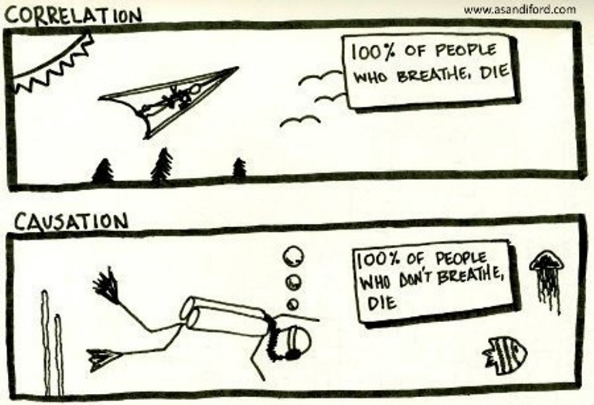
Causal Inference



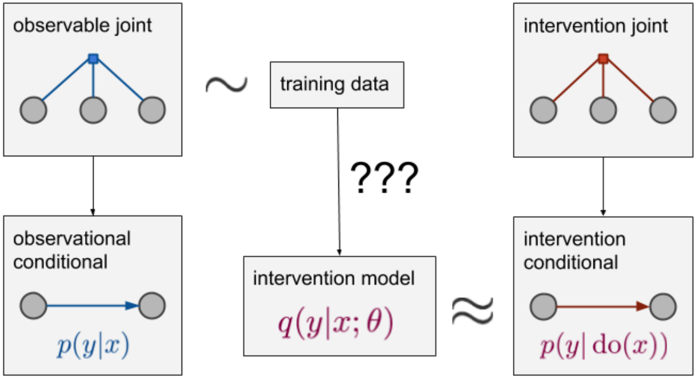
Causal Inference

Causal inference can be categorized into three different stages:

- Association
- Intervention
- Counterfactuals



Causal Inference



4

[4] *ML beyond Curve Fitting: An Intro to Causal Inference and do-Calculus*. https://www.inference.vc/untitled/.

Relevance to STAI

- Explainable AI helps us trust deep learning models, in any high stack decision making problems.

Relevance to STAI

- Explainable AI helps us trust deep learning models, in any high stack decision making problems.
- Causal explainability helps us to determine true cause and effect in the decision making process.

Main Contributions in the paper

Aim: To generate causal post-hoc explanations for any deep learning classifiers.

Key Contributions

Main Contributions in the paper

Aim: To generate causal post-hoc explanations for any deep learning classifiers.

Key Contributions

- Design of new conceptual framework.

Main Contributions in the paper

Aim: To generate causal post-hoc explanations for any deep learning classifiers.

Key Contributions

- Design of new conceptual framework.
- Regularization function to disentangle latent features into two groups.

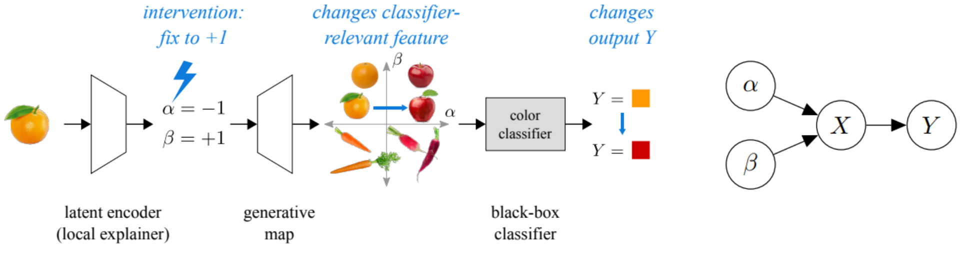
Main Contributions in the paper

Aim: To generate causal post-hoc explanations for any deep learning classifiers.

Key Contributions

- Design of new conceptual framework.
- Regularization function to disentangle latent features into two groups.
- Controlled experimentation.

Methodology



(a) (b)

Methodology

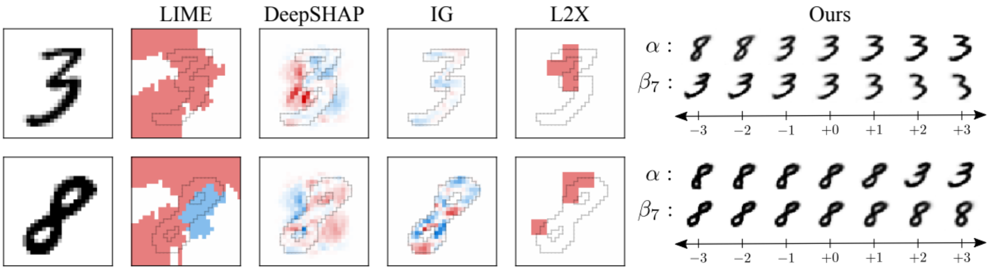
$$\arg\max_{g \in G} \quad \mathcal{C}(\alpha, Y) + \lambda \cdot \mathcal{D}\left(p(g(\alpha, \beta)), p(X)\right)$$

Proposition 2 (Information flow in our DAG)**.** *The information flow from α to Y in the DAG of Figure 1(b) coincides with the mutual information between α and Y. That is, $I(\alpha \to Y) = I(\alpha; Y)$, where mutual information is defined as $I(\alpha; Y) = \mathbb{E}_{\alpha,Y}\left[\log \frac{p(\alpha,Y)}{p(\alpha)p(Y)}\right]$.*

Algorithm 1 Principled procedure for selecting (K, L, λ).

1: Initialize $K, L, \lambda = 0$. Optimizing only $\mathcal{D}$, increase L until objective plateaus.
2: **repeat** increment K and decrement L. Increase λ until $\mathcal{D}$ approaches value from Step 1.
3: **until** $\mathcal{C}$ reaches plateau. Use (K, L, λ) from immediately before plateau was reached.

Results



Limitations

- Limited experimentation.

Limitations

- Limited experimentation.
- Method only works against linear generative class models.

Limitations

- Limited experimentation.
- Method only works against linear generative class models.
- It's hard to associate explanations to the classifier.

Possible future directions

- Incorporate classifiers influence to a greater extent in generating explanations.

[5] Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. "Deep structural causal models for tractable counterfactual inference". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 857–869.

Possible future directions

- Incorporate classifiers influence to a greater extent in generating explanations.
- Explore the possibility of non-linear generative models.

[5] Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. "Deep structural causal models for tractable counterfactual inference". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 857–869.

Possible future directions

- Incorporate classifiers influence to a greater extent in generating explanations.
- Explore the possibility of non-linear generative models.
- To incorporate an idea of DSCM[5] in the framework.

[5] Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. "Deep structural causal models for tractable counterfactual inference". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 857–869.

Questions?

Thank You!