Class Specific Interpretability in CNN Using Causal Analysis

Ankit Yadu[1], Suhas P K[1], Neelam Sinha[2]

[1]Samsung R&D Institute India, Bangalore (SRIB)

[2]International Institute of Information Technology, Bangalore (IIITB)

Contents

TREATMENT EFFECT ESTIMATION USING INVARIANT RISK MINIMIZATION

Abhin Shah,[†,⋆] Kartik Ahuja,[†] Karthikeyan Shanmugam,[†] Dennis Wei,[†]
Kush R. Varshney,[†] and Amit Dhurandhar[†]

[†]IBM Research, *Massachusetts Institute of Technology

ABSTRACT

Inferring causal individual treatment effect (ITE) from observational data is a challenging problem whose difficulty is exacerbated by the presence of treatment assignment bias. In this work, we propose a new way to estimate the ITE using the domain generalization framework of invariant risk minimization (IRM). IRM uses data from multiple domains, learns predictors that do not exploit spurious domain-dependent factors, and generalizes better to unseen domains. We propose an IRM-based ITE estimator aimed at tackling treatment assignment bias when there is little support overlap between the control group and the treatment group. We accomplish this by creating *diversity*: given a single dataset, we split the data into multiple domains artificially. These diverse domains are then exploited by IRM to more effectively generalize regression-based models to data regions that lack support overlap. We show gains over classical regression approaches to ITE estimation in settings when support mismatch is more pronounced.

***Index Terms*—** Causal inference, individual treatment effect estimation, invariant risk minimization

Table 1: A typical observational record from a hospital

Patient	Age	Blood Pressure	Drug	Blood sugar
A	22	145/95	0	Low
B	26	135/80	0	Low
C	58	130/70	1	Low
D	50	145/80	1	High
E	24	150/85	1	Low

from classical supervised learning because we never observe the ITE in our training data. For example, in Table 1 we do not observe the blood sugar under *the treatment* for patients in the *control group* and the blood sugar under *the control* for patients in the *treatment group*.

Unlike RCTs, observational data is often prone to treatment assignment bias [9]. For instance, patients receiving drug '0' may have a higher natural tendency (due to their age) to have low blood sugar than patients receiving drug '1'. In other words, sub-populations receiving different treatments can have very different distributions, and a traditional supervised learning model trained to predict the effect of treatment would fail to generalize well to the entire population. This

Causal Discovery from Soft Interventions with Unknown Targets: Characterization and Learning

Amin Jaber
Department of Computer Science
Purdue University, USA
jaber0@purdue.edu

Murat Kocaoglu
MIT-IBM Watson AI Lab
IBM Research MA, USA
murat@ibm.com

Karthikeyan Shanmugam
MIT-IBM Watson AI Lab
IBM Research NY, USA
karthikeyan.shanmugam2@ibm.com

Elias Bareinboim
Department of Computer Science
Columbia University, USA
eb@cs.columbia.edu

Abstract

One fundamental problem in the empirical sciences is of reconstructing the causal structure that underlies a phenomenon of interest through observation and experimentation. While there exists a plethora of methods capable of learning the equivalence class of causal structures that are compatible with observations, it is less well-understood how to systematically combine observations and experiments to reconstruct the underlying structure. In this paper, we investigate the task of structural learning in non-Markovian systems (i.e., when latent variables affect more than one observable) from a combination of observational and soft experimental data when the interventional targets are unknown. Using causal invariances found across the collection of observational and interventional distributions (not only conditional independences), we define a property called Ψ-Markov that connects these distributions to a pair consisting of (1) a causal graph $\mathcal{D}$ and (2) a set of interventional targets $\mathcal{I}$. Building on this property, our main contributions are two-fold: First, we provide a graphical characterization that allows one to test whether two causal graphs with possibly different sets of interventional targets belong to the same Ψ-Markov equivalence class. Second, we develop an algorithm capable of harnessing the collection of data to learn the corresponding equivalence class. We then prove that this algorithm is sound and complete, in the sense that it is the most informative in the sample limit, i.e., it discovers as many tails and arrowheads as can be oriented within a Ψ-Markov equivalence class.

CANDLE: An Image Dataset for Causal Analysis in Disentangled Representations

Abbavaram Gowtham Reddy
IIT Hyderabad, India
cs19resch11002@iith.ac.in

Benin Godfrey L
IIT Hyderabad, India
benin.godfrey@cse.iith.ac.in

Vineeth N Balasubramanian
IIT Hyderabad, India
vineethnb@iith.ac.in

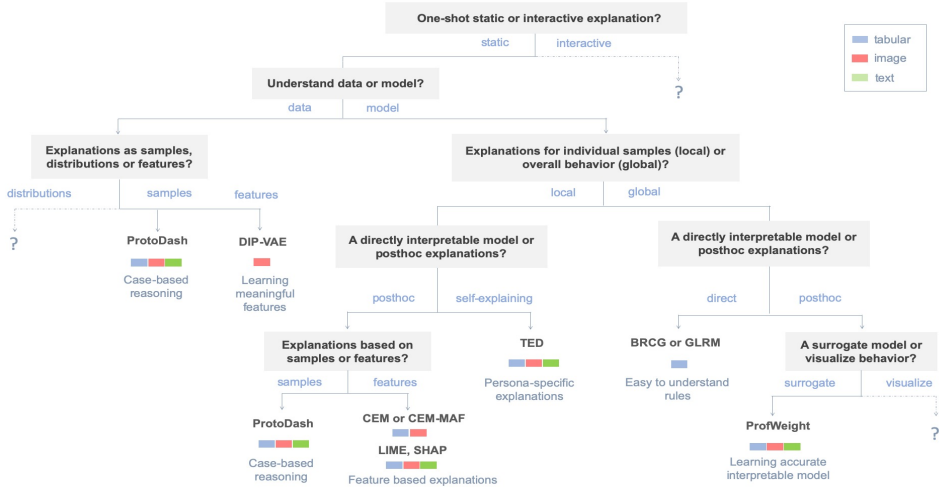
Abstract

Confounding effects are inevitable in real-world observations. It is useful to know how the data looks like without confounding. Coming up with methods that identify and remove confounding effects are important for various downstream tasks like classification, counterfactual data augmentation, etc. We develop an image dataset for Causal ANalysis in DisentangLed rEpresentations(CANDLE). We also propose two metrics to measure the level of disentanglement achieved by any model under confounding effects. We empirically analyze the disentanglement capabilities of existing methods on dSprites and CANDLE datasets.

Figure 1: Sample images from CANDLE.

current models either assume confounding is not present, or ignore it even if it is. We encourage models that consider confounding by creating a dataset with both observed and

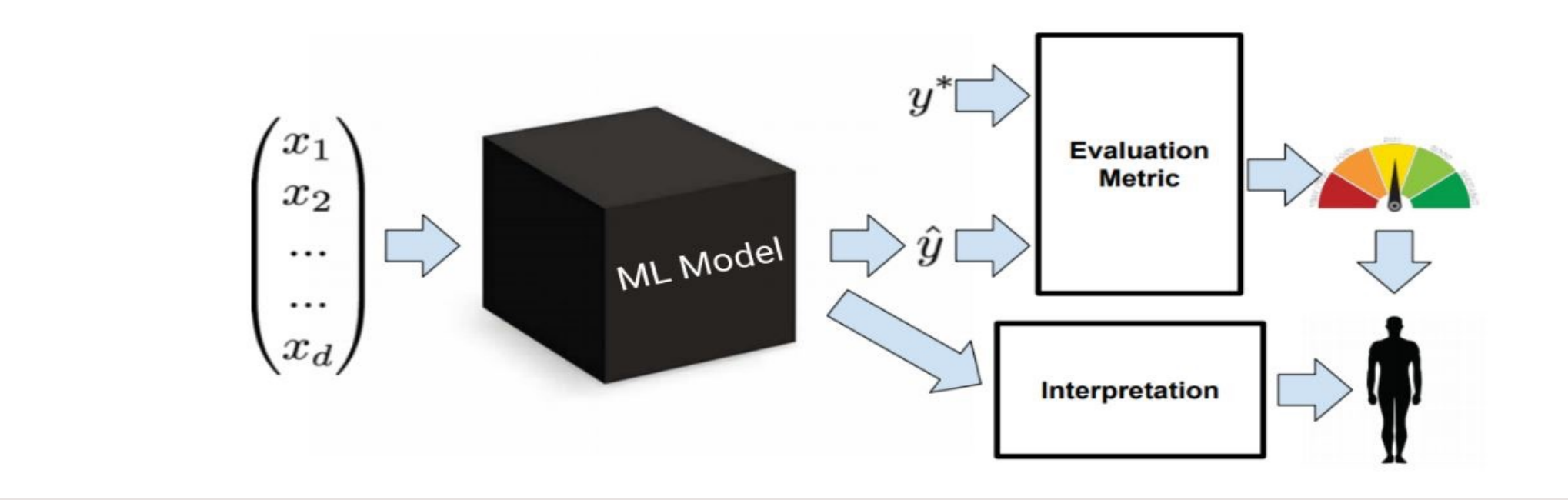
AI Explainability 360: An Extensible Toolkit for Understanding Data and Machine Learning Models*



*Vijay Arya and Rachel K. E. Bellamy and Pin-Yu Chen and Amit Dhurandhar and Michael Hind and Samuel C. Hoffman and Stephanie Houde and Q. Vera Liao and Ronny Luss and Aleksandra MojsiloviÄ‡ and Sami Mourad and Pablo Pedemonte and Ramya Raghavendra and John T. Richards and Prasanna Sattigeri and Karthikeyan Shanmugam and Moninder Singh and Kush R. Varshney and Dennis Wei and Yunfeng Zhang, JMLR 2020.

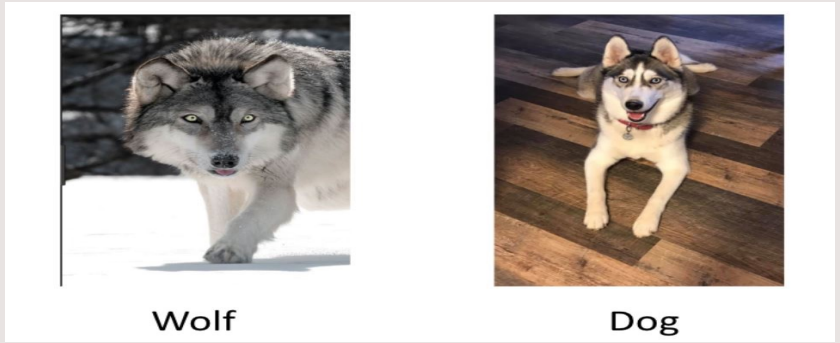
1. Introduction

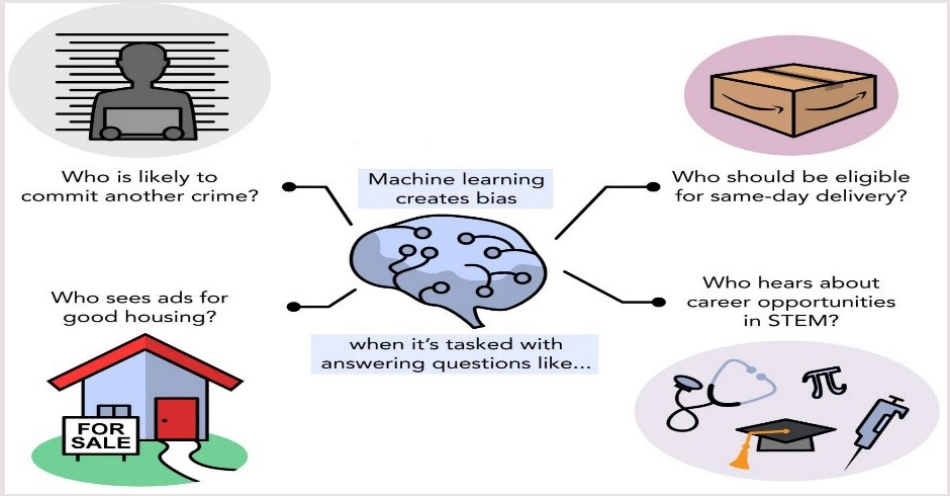
Interpretation of a machine learning model is the process wherein we try to understand the predictions of a machine learning model.

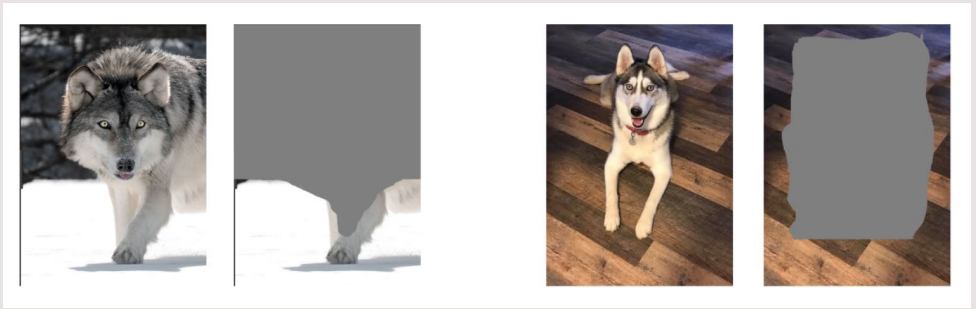


1. Introduction (cont...)

Why do we need interpretable Machine learning ?







- Fairness
- Features learnt & model debugging
- Regulations

Contents

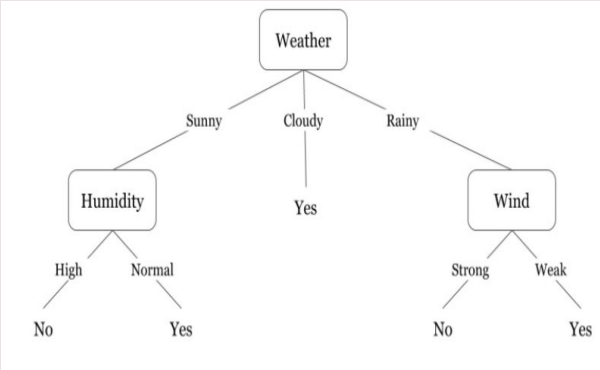
2. Interpretability Methods

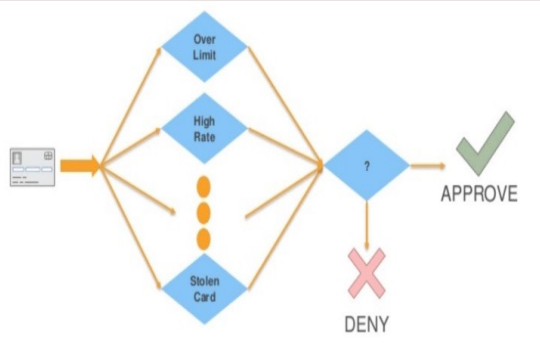
- Traditional Interpretable methods
    - Inherently interpretable models.
    - Post-hoc interpretable models.

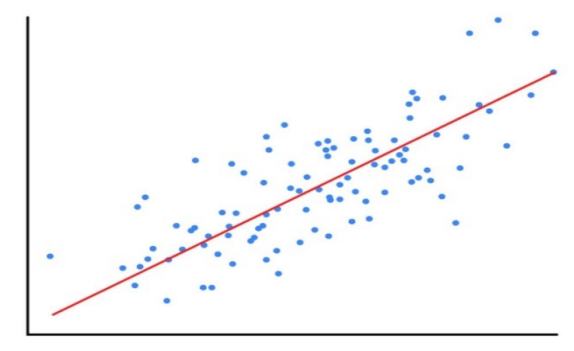
- Causal Interpretability

2. Interpretability Methods

Inherently interpretable models







- Decision Trees
- Rule based models
- Linear Regression
- Attention Networks

- Disentangled learning
  - PCA
  - Spectrum Analysis

2. Interpretability Methods (cont...)

Post-hoc Interpretability

- Local Interpretable Model-Agnostic Explanations

  M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016

- Saliency Maps

  M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In European conference on computer vision, 2014

- Example Based Explanations

  B. Kim, R. Khanna, and O. O. Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In Advances in Neural Information Processing Systems, 2016

- Feature Visualization

  D. Erhan, Y. Bengio, A. Courville, and P. Vincent. Visualizing higher-layer features of a deep network. University of Montreal

- Explaining by Base Interpretable Model

  M. Craven and J. W. Shavlik. Extracting tree structured representations of trained networks. In Advances in neural information processing systems,2016

Contents

3. Causal Inference

Structural Causal Model (SCM)

- An SCM is a 4-tuple variable M(X, U, f, P_u)

  X : finite set of endogenous variables

  U : finite set of exogenous variables

  f : set of function $\{f_1, f_2, \dots, f_n\}$, represents casual mechanism such that,

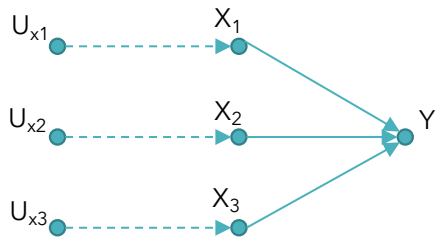
  $x_i = f_i(Pa(x_i), u_i) \qquad \forall x_i \in X$

  $Pa(x_i) \subseteq (X \setminus \{x_i\} \cup U$

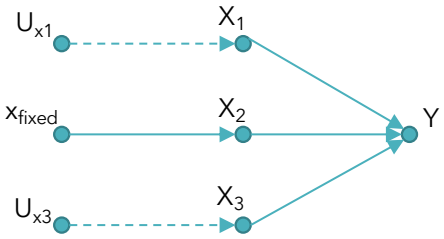
- Original SCM is represented by below distribution function
  - $Y = P(y \mid x_1, x_2, x_3)$

- Interventional SCM is represented as:
  - $Y = P(y \mid x_1, do(X_2 = x_{fixed}), x_3)$

Original SCM



Interventional SCM



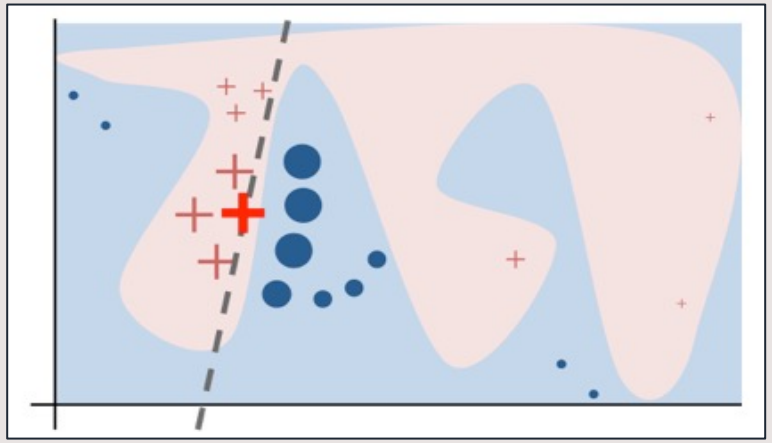
Contents

4. Prior Work (cont....)

[1]Sundararajan et al. (2017). *ICML*
[2] Kindermans et al. (2017). *NIPS*
[3] Ribeiro et al. (2016). *ACM SIGKDD*

- Axioms of Attribution
  - Sensitivity[1]
    - $\forall x_i \in X, \quad f(x_i) \neq f(x_i^{baseline}) \implies A_i^f(x) \neq 0$
  - Implementation Invariance[1]
    - $\forall x_i \in X, \quad f_1(x_i) = f_2(x_i) \implies f_1 \equiv f_2$
  - Completeness[1]
    - $f(x) = \sum_i A_i^f(x)$
  - Symmetry Preserving[1]
    - $f(x, y) = f(y, x) \implies A^f(x) = A^f(y)$
  - Input Invariance[2]

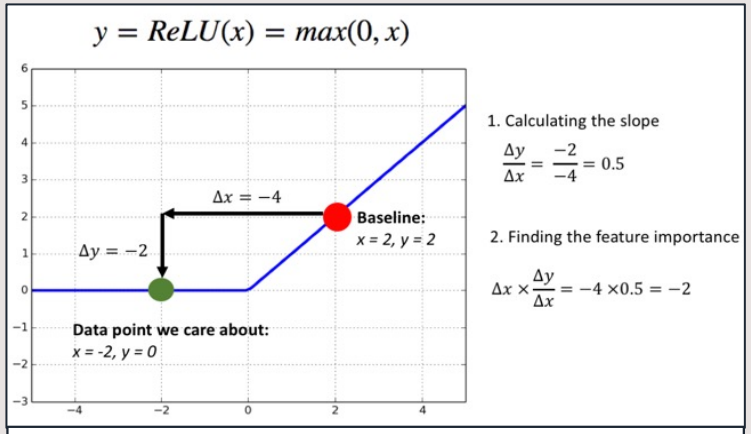
- LIME[3] - Locally Interpretable Model-Agnostic Explanations
  - Perturbation-based method
  - Generates an explanation for an instance with a surrogate model
  - Each instance has a different interpretable model for explanation.



4. Prior Work (cont...)

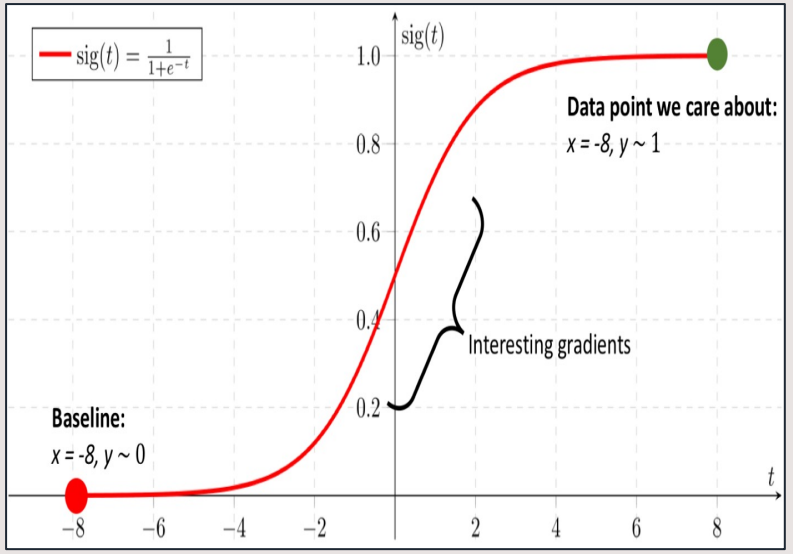
[1] Shrikumar et al. (2017). *ICML*
[2] Sundararajan et al. (2017). *ICML*

- DeepLIFT[1] – Deep Learning Important Features
  - Assigns importance score based on difference from reference policy (baseline)
  - Addresses model saturation and thresholding problem.
  - Approximates instantaneous gradients to explain the change in slope w.r.t baseline



Pure gradient based approach would have assigned an importance score of 1 for (x,y) = (-2, 0) with baseline as (2, 2).

- Integrated Gradients[2] : Integral of gradients averaged along a path from baseline to a particular feature
  - $IG_i(x) = (x_i - x_i') \int_\alpha^1 \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} \partial\alpha$



4. Prior Work (cont...)

Probabilistic based causal approach[1]

[1]Chattopadhyay et al. (2019). *ICML*

- Computes causal effect of each feature on the output.
- Considers neural network model as SCM to perform causal estimation.
- Our work is inspired by this prior work on using ACE in neural network.

Contents

5. Proposed Method

Absolute area of ACE (A-ACE)

- $ACE^{y}_{do(x_i \in (\alpha_1, \alpha_2))} = \int_{\alpha_1}^{\alpha_2} E[y \mid do(xi = \alpha)] - E[y \mid do(xi = x')]\, \partial\alpha$

- $\text{A-ACE} = \int_{\alpha_1}^{\alpha_2} |ACE|\, \partial\alpha$

Here –

- y is Average Causal Effect (ACE) estimate w.r.t feature α
- α_1 and α_2 are minimum and maximum interventional values for feature α

5. Proposed Method

Derivation

- A-ACE

$$A - ACE^{y}_{do(x_i \in (\alpha_1, \alpha_2))} = \int_{\alpha_1}^{\alpha_2} |E[y \mid do(x_i = \alpha)] - E[y \mid do(x_i = x')]| \, \partial\alpha$$

$$E[y \mid do(x_i = \alpha)] = \int yp(y \mid do(x_i = \alpha) \, dy$$

- Employs Taylor series expansion to compute the interventional expectation

$$E[f'_y(l_1) \mid do(x_i = \alpha)] \approx f'_y(\mu) + \frac{1}{2} Tr(\nabla^2 f'_y(\mu) \, E[(l_1 - \mu)(l_1 - \mu)^T) \mid do(x_i = \alpha)]$$

$$\mu = [\mu_1, \mu_2, \dots, \mu_k]^T$$

$$\mu_j = E[x_j \mid do(x_i = \alpha)]$$

$$l_1 = [x_1, x_2, \dots, x_k]$$

5. Proposed Method (cont...)

Limitations of ACE for interpretability

- Causal analysis is possible through intervention on the model.

- ACE doesn't exploit the variation with changing levels of interventional values.

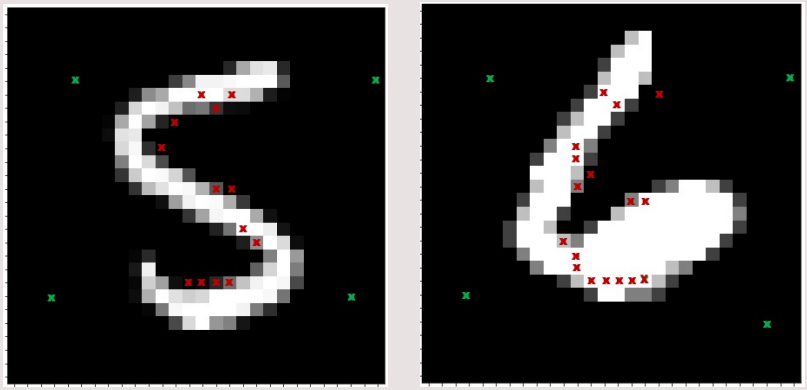
5. Proposed Method (cont...)

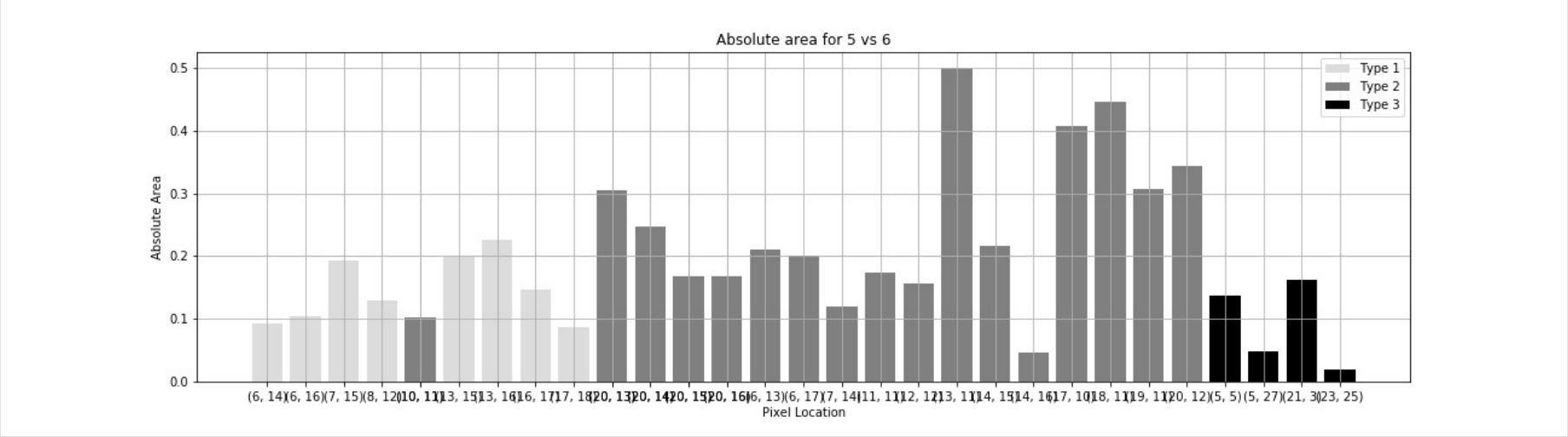
Significance of the method

- A-ACE improves the interpretability quantification by exploiting the magnitude of causal effect.
- Interpretability is possible only by localizing discriminative regions.
- In our experiments we show that the behaviour at the discriminative regions can be captured using the proposed method (A-ACE).
    - We have incorporated three types of regions for MNIST – pixels common to both the classes, distinguishing pixels, and background pixels.
    - For ILSVRC dataset we have included regions related to $class_i$ and background pixels to interpret the most relevant regions of class object learnt by the model to classify for $class_i$

Contents

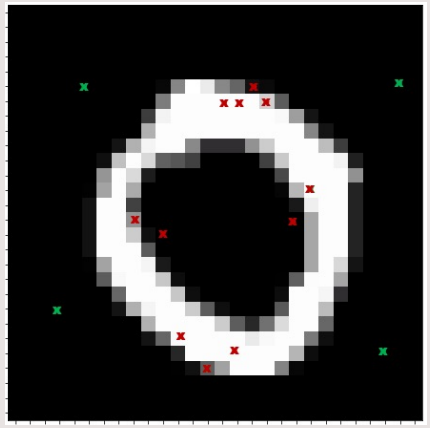
6. MNIST digits and results | 5 vs 6

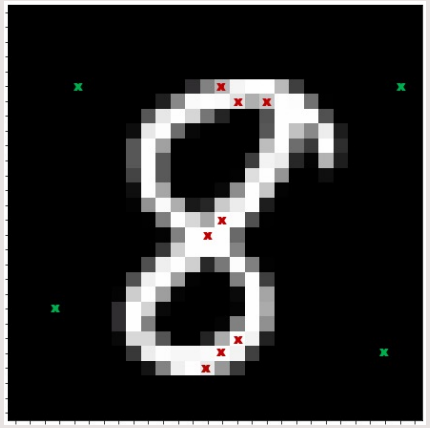


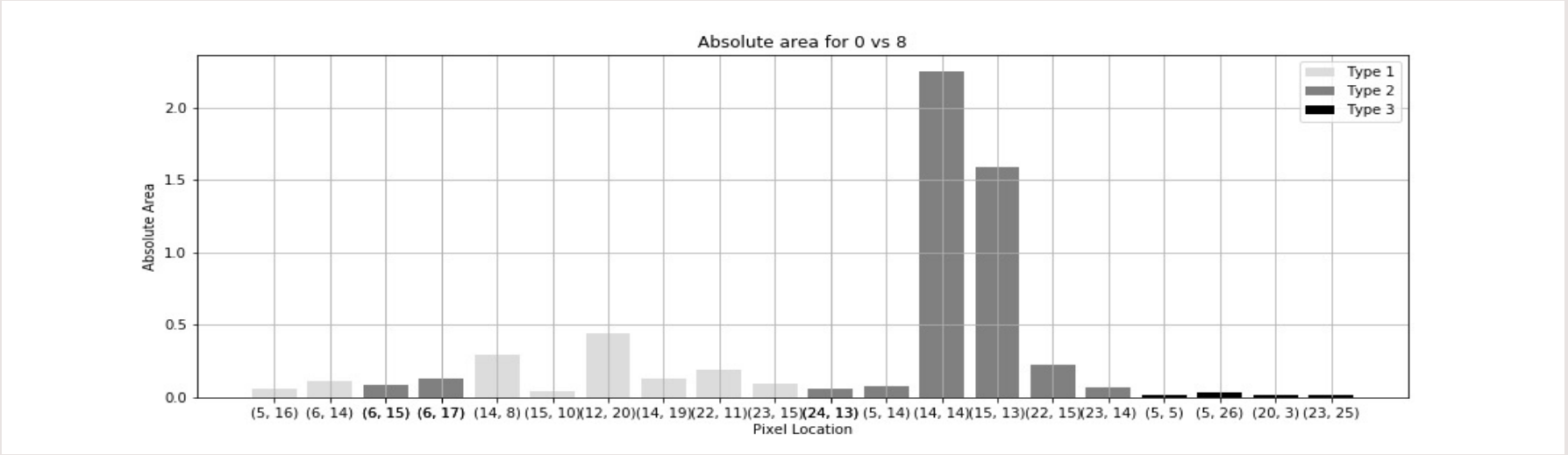


6. MNIST digits and results

0 vs 8

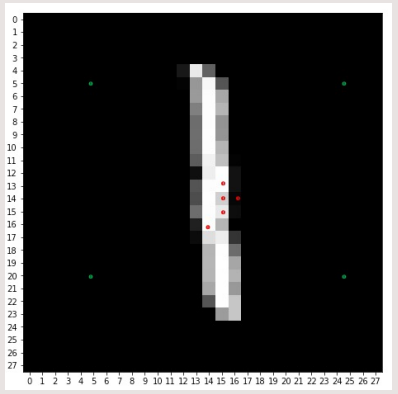


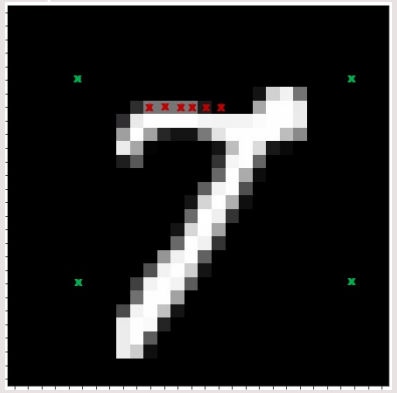


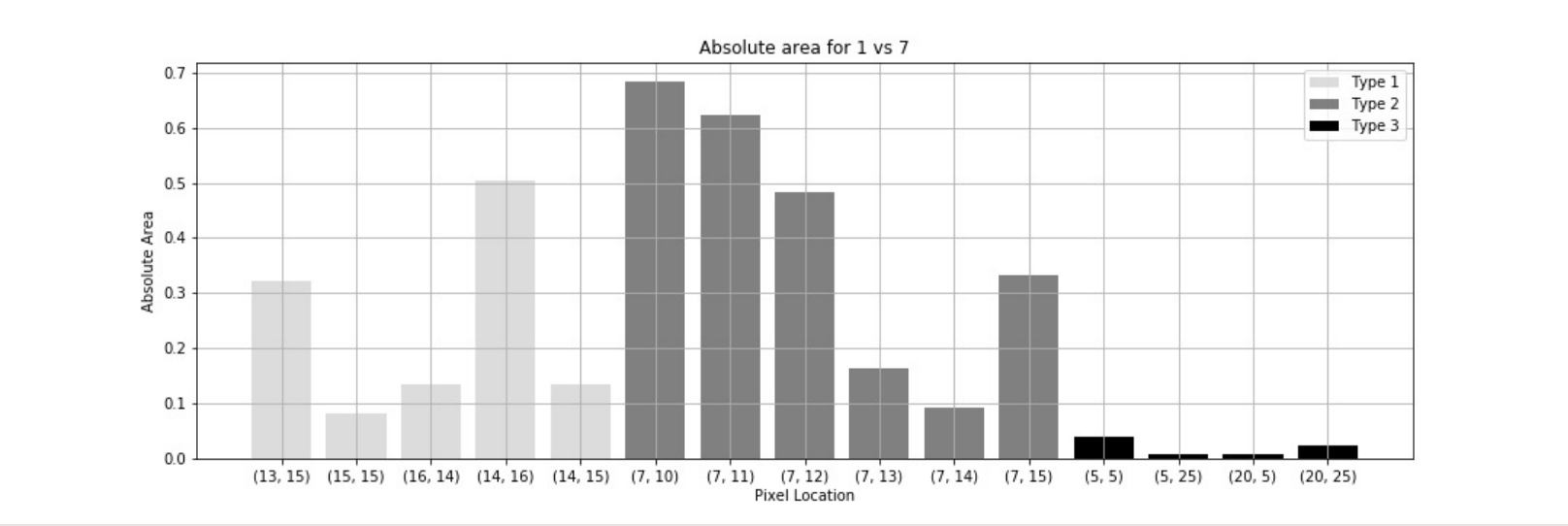


6. MNIST digits and results

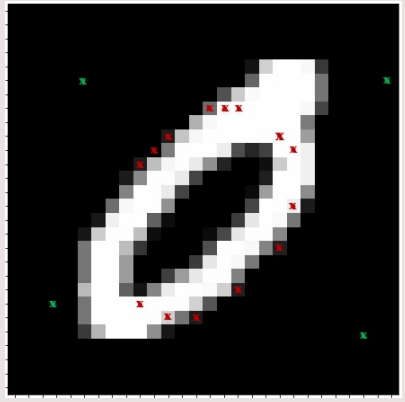
1 vs 7

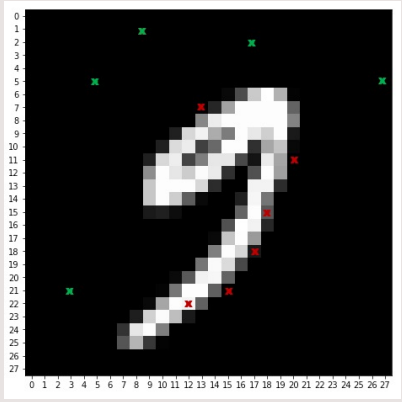


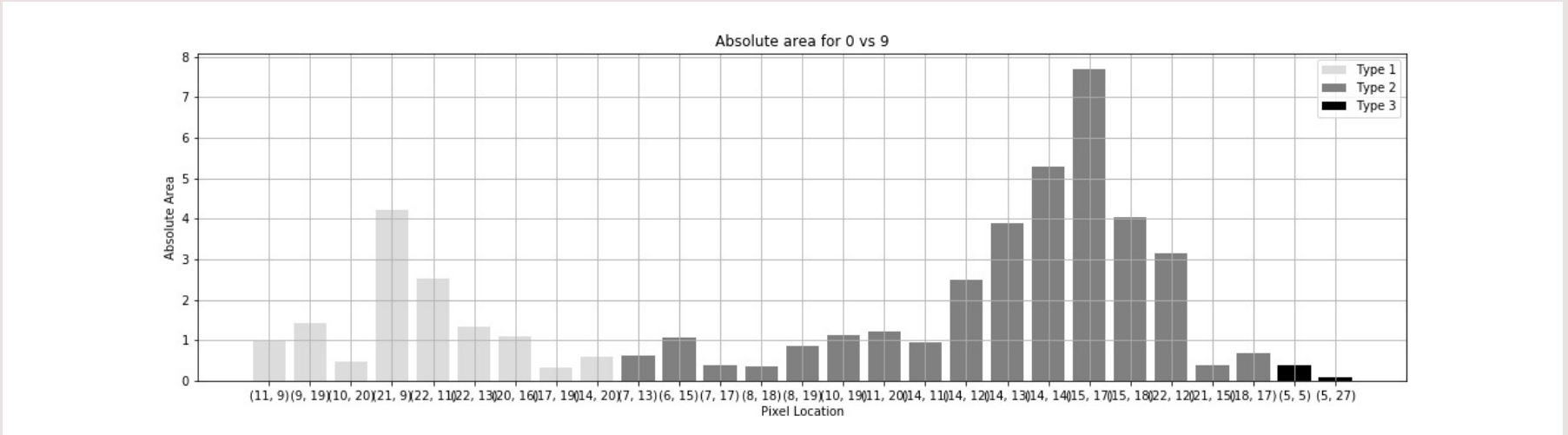




6. MNIST digits and results | 0 vs 9



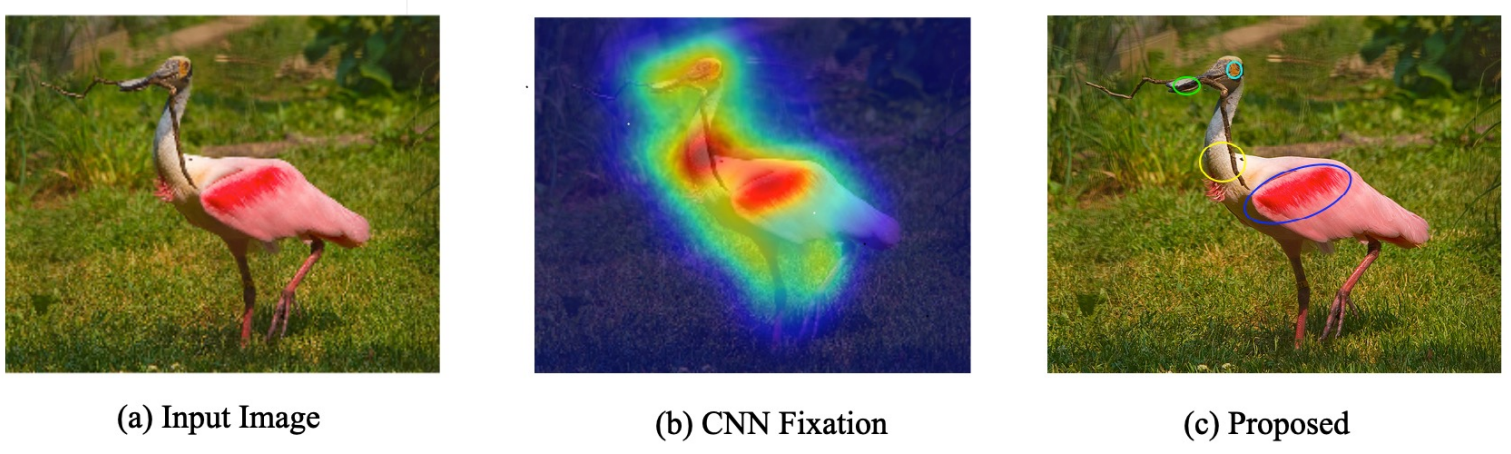




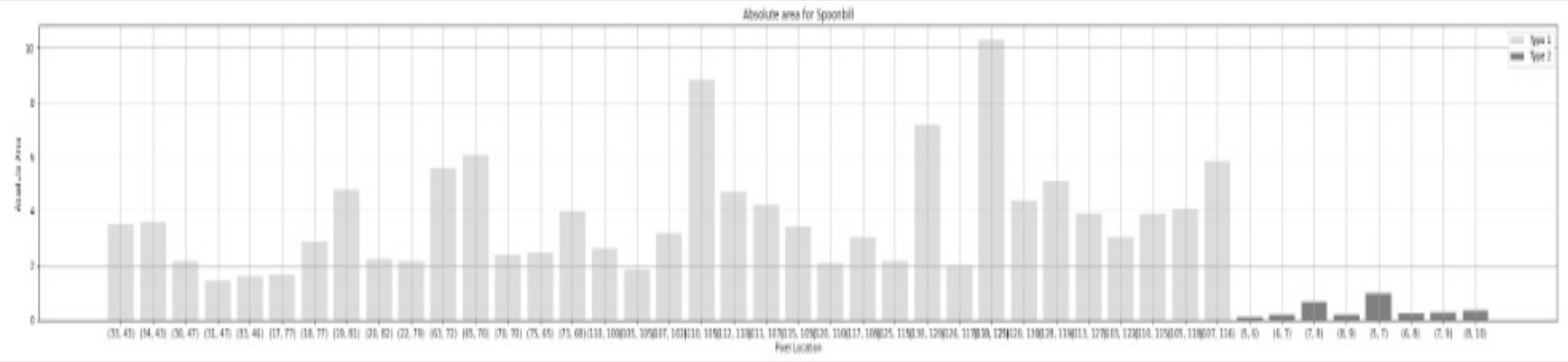
Contents

7. Comparison with SOTA (CNN Fixation)

ResNet-101 model trained on ILSVRC dataset is used. The sample image is taken from validation dataset



(a) Input Image (b) CNN Fixation (c) Proposed



Contents

8. Limitations

- A-ACE estimation requires full interventional range for each input feature.

- Computational time of A-ACE is directly proportional to the model's complexity (number of layers, size of layers etc.)

Contents

9. Conclusion

- Proposed method can be easily extended to complex models by appropriately changing the shape of input data, gradients and hessian matrix, mean and covariance matrix.

- Classification problem reformulated as a binary classification problem. We consistently find peak at distinguishing pixels to be at least 33% higher than other for MNIST data and 10x higher in ILSVRC data.

- A-ACE exploits the magnitude of the causal effect irrespective of the direction. This leads to improved quantification of interpretability