CLASS SPECIFIC INTERPRETABILITY IN CNN USING CAUSAL ANALYSIS

Ankit Yadu* *Suhas P K** *Neelam Sinha*†
*Samsung R&D India, Bangalore $\{ankit.yadu, suhas.pk\}$@samsung.com
†International Institute of Information Technology, Bangalore $\{neelam.sinha\}$@iiitb.ac.in

ABSTRACT

A singular problem that mars the wide applicability of machine learning (ML) models is the lack of generalizability and interpretability. The ML community is increasingly working on bridging this gap. Prominent among them are methods that study causal significance of features, with techniques such as Average Causal Effect (ACE). In this paper, our objective is to utilize the causal analysis framework to measure the significance level of the features in binary classification task. Towards this, we propose a novel ACE-based metric called "Absolute area under ACE (A-ACE)" which computes the area of the absolute value of the ACE across different permissible levels of intervention. The performance of the proposed metric is illustrated on MNIST data set (~42000 images) by considering pair-wise binary classification problem. The computed metric values are found to be higher (peak performance of 50% higher than others) at precisely those locations that human intuition would mark as distinguishing regions. The method helps to capture the quantifiable metric which represents the distinction between the classes learnt by the model. This metric aids in visual explanation of the model's prediction and thus, makes the model more trustworthy.

***Index Terms*—** Causal Inference, Interpretability, CNN, Explainability, Machine Learning

1. INTRODUCTION

Machine learning models have enormous potential in solving sophisticated problems in variety of domains be it object detection [8], autonomous driving [1], DNA sequence generation [17], speech recognition [24] and language processing [5]. These models are capable of learning complex representations. However, human interpretability of these models has been very challenging due to the Black Box nature of it, thus making these models untrustworthy to be used in critical scenarios like healthcare applications as in disease discovery and diagnosis [11], drug discovery [9], autonomous driving [1] etc. Trusting these models in critical application requires us to be cognizant about the pertinent features and their effectiveness, that the model has learnt. These models would have been validated only on the perceived scenario and lacks to accommodate the unseen scenarios - drifting distribution. Hence there is a high need to strengthen the modelling methods for meaningful predictions from such models. The emerging field in machine learning, Explainable AI [6], aims to address this problem of discovering the Black Box decisions in deep neural networks.

Various Methods have been proposed to overcome the problem of interpretability, each having its own perspective towards Explainable AI [6]. Arguably, the machine learning model takes raw input features and learns the mapping between input and output. The learnt mapping doesn't help to understand how changing a feature would affect model's prediction. The existing method [28] does an exhaustive search on the feature space, where the permutation of the features each with its own range of values are used to assign the feature with an importance score. This method exponentially increases the computational time and thus, is unscalable. Hence, a much simpler approach, causal inference method has been proposed in this paper which helps us to analyse the causal relation of each feature and has been validated with the feature's sensitivity.

We have tried to address the problem of interpretability by proposing perturbation-based framework for causal inference. The objective of the proposed work is to formulate the digit classification as a two-class classification problem and understand the model's prediction on the distinctive region of the two classes. Achieving this would help in improving the reliability of the model, thus making it better suitable for deploying in critical scenarios like healthcare applications, autonomous driving etc.

2. PRIOR WORK AND MOTIVATION

The advancement of Deep Neural Networks has motivated many researchers to investigate feature attribution. Some efforts in this regard DeepLIFT [18], Layerwise Relevance Propagation (LRP) [21], Locally Interpretable Model (LIME) [20], Causal Framwork for sequence to sequence to models [15] and Probabilistic based Causal approach [7].

For the analysis of attribution-based approach, [19], contributed axioms that could be used to understand model's prediction. [19] explains fundamental axioms for attribution-based methods. The explainable method must satisfy Sensitivity, Implementation Invariance, Completeness, Symmetry Preserving, and Integrated Gradients to make the prediction transparent and trustworthy.

A perturbation-based method, LIME framework, generates local explanation for the decision function. A particular instance is perturbed to get the original model's prediction and based on the output, it attempts to explain the prediction with an interpretable model like – linear model, decision tree etc. The explanation would depend on the particular instance being explained; hence the framework is a local model, generating explanations around an instance. On the other hand, in gradient based approach, one would take the gradient of model's output w.r.t to the feature to assign relevance score. This method could not hold (i) because feature importance is relative to the model and neural-network model fails to reveal how a feature changed the prediction relative to the model not seeing the feature at all, *baseline*, and (ii) the gradient based approach breaks for non-linear models. An ideal baseline is a point on the decision boundary where the predictions are neutral, [19]. To overcome the limitation of gradient based approach, DeepLIFT, takes the model's gradient to explain how output changes, based on change in input from its baseline. [18] have approximated the gradients for explaining the prediction w.r.t baseline, which violates implementation invariance as in [19]

Probabilistic based casual approach, [7], a recent work contributed on a causal based attribution method for neural network models which helps to compute the causal effect of each feature on the output. It relies on the work of considering DNN models as an SCM, [16]. Additionally, it includes efficient algorithm to compute the causal effect.The objective of [7] is complementary to the reported study. The proposed approach studies the causal impact of the input features in the image space, by proposing a novel ACE-based metric but the authors in [7] study the impact in the latent space - rotation, shape change, on MNIST data. Another related work [2] analyzes the latent-space representation to examine the causal relations using ACE on MNIST dataset.

CNN fixations [13] are also used to interpret class differences. However, it can be used to explain inter-class difference only after obtaining perfect classification. On the other hand, causal relationships can be explored without any such dependencies. In order to have a method both causal and interpretable, it is necessary to provide an explanation of how the results have been achieved [4]. There are certain benchmark dataset which are utilized to evalute the interpretability like ImageNet (ILSVRC) [22], MNIST [3]. However, extending the same on other dataset which does not comprises of human-grounded truth is difficult to verify without formal methods. In the proposed work we attempt to incorporate sensitivity analysis on the trained model to verify the causal inference results.

In order to elaborate our work, it is necessary to highlight the key principles of causal inference which we have employed as a basis for CNN interpretation.

2.1. CAUSAL INFERENCE

Definition 1. *(Structual Causal Model [25], [27]). A 4-tuple variable M(X, U, f, P_u) where X is a finite set of endogenous variables, usually the observable variables, U denotes a finite set of exogenous variables which usually account for unobserved or noise variables, f is a set of function $\{f_1, f_2, ..., f_n\}$ where each function represents a causal mechanism such that $\forall x_i \in X$, x_i = f_i(Pa(x_i), u_i) and Pa(x_i) is a subset of (X x_i) $\cup$ U and P_u is a probability distribution over U is called a Structural Causal Model (SCM).*

Definition 2. *(Causal Bayseian Network). A directed graphical model G(V, E) is used to represent an SCM M(X, U, f, P_u). V is the set of endogenous variables X and E denotes the causal mechanisms. This indicates for each causal mechanism x_i = f_i(Pa(x_i), u_i), there exists a directed edge from each node in the parent set Pa(x_i) to x_i. The entire graph representing this SCM is called a Causal BayesianNetwork (CBN)*

Definition 3. *(Average Causale Effect [25], [27]). The Average Causal Effect(ACE) of a binary variable x (treatment) on another random variable (outcome) is define as*

$$ACE = \mathbb{E}[y|do(x=1)] - \mathbb{E}[y|do(x=0)] \tag{1}$$

Where do(.) [23] operator denotes the corresponding interventional distribution defined by the SCM or CBN.

Extending eqn. (1), in case of a continuous random variable, we would need a definite baseline, $\widehat{x_i}$, for each input feature, and hence, the average causal effect of x_i, after being intervened with α, on y, [7], is captured in eqn. (2)

$$ACE^{y}_{do(x_i=\alpha)} = \mathbb{E}[y|do(x_i=\alpha] - \mathbb{E}[y|do(x_i=\widehat{x_i})] \tag{2}$$

Furthermore, the interventional expectation, expected value of y when x_i is intervened with α, can be evaluated by eqn. (3)

$$\mathbb{E}[y|do(x_i=\alpha)] = \int_y yp(y|do(x_i=\alpha))dy \tag{3}$$

3. PROPOSED METHOD

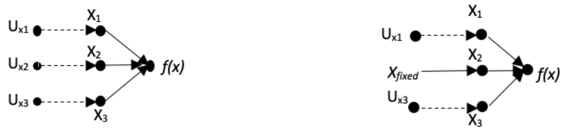
The proposed metric is based on a popular measure used in causal analysis called ACE.

The rules of causal inference aid in estimating the Average Causal Effect (ACE) of an individual feature on the model function, $f(x)$, by intervening the particular input feature node with a distribution function as in fig. 1(a). The illustration in fig. 1(b) indicates one of the input feature, x_2, has been intervened by clamping it to x_0 (interventional value).

$$f(x) = p(y|x_1, x_2, x_3) \tag{4a}$$

$$f(x) = p(y|x_1, do(X_2 = x_0), x_3) \tag{4b}$$

Eqn. 4a represents the conditional probability of y when x_1, x_2, and x_3 are observed. Eqn. 4b represents the interventional probability of y when x_1 and x_3 are observed, and x_2 is intervened, by clamping its value to x_{fixed}, which can be evaluated using inference rules of ***do***-calculus [23].



(a) Original SCM (b) Interventional SCM

Fig. 1: SCM. Unobserved (exogenous) variables U_i are connected by dashed arrows. Each input feature X_i has a corresponding unmeasurable influencing variable, U_i. 1(a) illustrates original SCM encoding the presence of causal influence of X on $f(x)$ by X $\rightarrow$ Y and absence of any influence of $f(x)$ on X. 1(b) depicts, one of the feature X_2 intervened with a random value x_{fixed}, and hence called an interventional SCM.

An intervention on a particular input feature with a continuous domain would, in effect, change the distribution of the output function. The analysis of this interventional distribution would help us to understand the causal attribution of the particular feature like - how the causal attribution changes over the intervention and what is the strength of the causation in a particular interventional range. The causal attribution is captured by estimting the ACE for each feature w.r.t output. The ACE plot depicts feature-wise average deviation of the causal attribution from the baseline for a range of interventional values.

3.1. Proposed Metric : Absolute Area Under ACE "A-ACE"

Although ACE is a very popular measure, existing studies do not utilize the pattern of variation of the values of ACE across the different levels of intervention. In most reported works significant features are identified based on the levels of ACE that they elicit. For a considered feature, the proposed metric integrates out the magnitude of ACE for all permissible levels of intervention and regards the obtained value as a measure of the causal contribution of that feature.

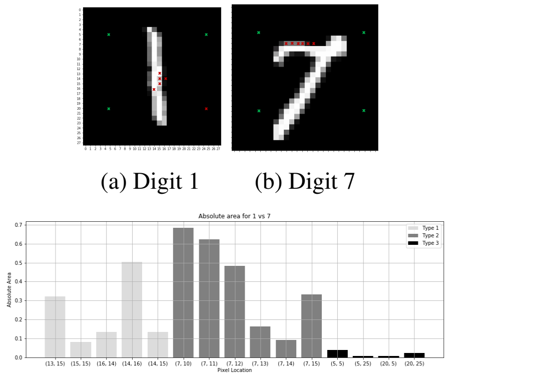
$$\text{A-ACE} = \int_{\alpha_1}^{\alpha_2} |y| d\alpha \tag{5}$$

Eqn. 5 represents area under absolute value of ACE, from α_1 to α_2. Here, y is the ACE estimate w.r.t feature α. α_1 and α_2 are the minimum and maximum interventional value of α, respectively. We are interested in these α values as they don't alter the model's prediction. These values are computed for each feature separately, i.e, as many features those many ACE plots.

4. EXPERIMENTS AND RESULTS

In our work, we have attempted to uncover the black-box nature of convolutional neural network-based model with the open dataset: MNIST [3]. In this work, we have employed CNN model with 2 convolutional layers and 2 dense layers and formulated digit classification as a two-class classification problem. The two classes are picked after thorough visual inspection in such a way that they have an overlapping region which we wanted the model to learn as depicted in figure 2 and 3.

We have used taylor series expansion to approximate the interventional expectation. The handpicked pixels from each digit are marked by red color and the background pixels are marked by green color in the sample digits in subfigures (a) and (b) of figure 2, 3, 4, and 5. The corresponding subfigure (c) in each figure 2, 3, 4, and 5 shows the absolute area of ACE for the handpicked pixels. Clearly, the set of pixels which distinguishes the two classes have the highest absolute area of ACE. This confirms that the CNN model has learnt the expected class pattern and also that the absolute area of ACE can be used as visual explanation to interpret the model's prediction. Illustration of absolute area of ACE in figure 2(c), 3(c), 4(c), 5(c) encodes type 1 pixels as skeleton of the digit, type 2 pixels as discriminative pixels among the classes and type 3 as background pixels.

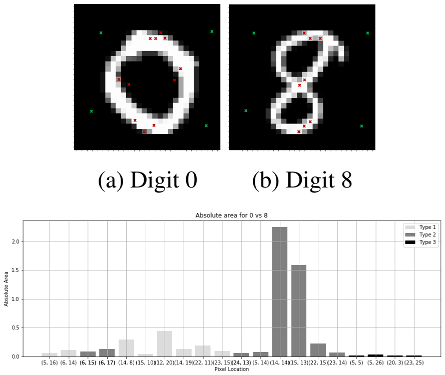


(a) Digit 1 (b) Digit 7

(c) Absolute area of ACE

Fig. 2: Sample digits (a) 1 and (b) 7. (c) Proposed Metric : Absolute area of ACE "A-ACE" for handpicked pixels.

As seen from the plots a simple thresholding could be used to pick out the highly significant pixels that are critical for discrimination between the two classes. As part of this study, the total area of ACE was also computed. This computation, however, did not lead to meaningful interpretation of the features. This can be attributed to the fact that the significance of a feature is directionless and does not depend on the sign of ACE value it elicits.



(a) Digit 0 (b) Digit 8

(c) Absolute area of ACE

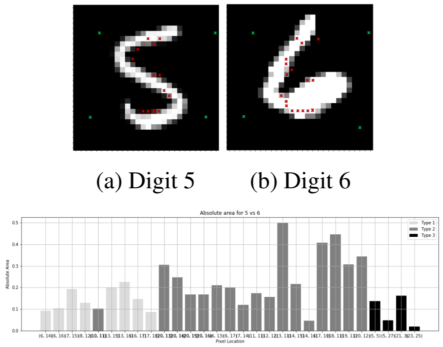
Fig. 3: Sample digits (a) 0 and (b) 8. (c) Proposed Metric : Absolute area of ACE "A-ACE" for handpicked pixels

5. DISCUSSION

CNN fixation is yet another approach to determine the features that significantly contribute to classification. However, fixations can be computed only after perfect classification. Besides computing fixation is computationally more expensive since it requires backtracking on the forward pass to locate the significant pixels. Also, several existing methods utilize ACE for evaluating causal attribution of the features. For eg [12] estimates ACE when high-dimensional co-variate information is available, however, such information is not required in our work. [14] reports a technique to estimate the ACE measure itself. The distinguishing aspect of the proposed metric compared to those in literature is that the pattern in variation of the magnitude of ACE is exploited as opposed to viewing values of ACE of distinguishing features as relatively significant for hand-picked interventional values. The experiments illustrate that the proposed metric A-ACE, which is the area under the magnitude of the ACE curve, serves well to quantify the causal attribution of a feature. The reported results form a pilot study of efficacy of the proposed metric. More elaborate experiments with complex, real-world images need to be carried out to establish the credibility of the metric.

6. CONCLUSION

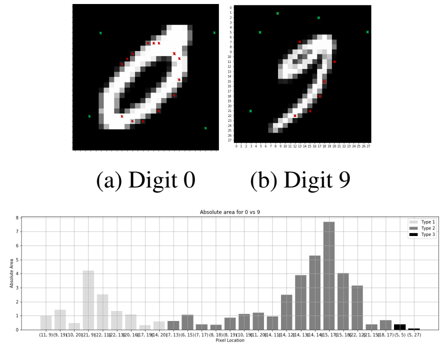
In this work, we have employed principles of causal inference to unwrap the pattern learnt by a CNN model using MNIST[3] data. The digit classification problem is reformulated as a two-class classification problem, whose results can be intuitively verified . A novel metric based on the popular ACE, called Absolute Area Under ACE curve, "A-ACE", has been proposed. This metric is computed by integrating out the absolute values of ACE over all permissible levels of intervention. Experiments on ~42000 images illustrate the performance of the proposed metric. The computed metric resulted in peak performance of 50% higher values compared to the relatively non-significant features. This metric could be used towards making models more trustworthy.



(a) Digit 5 (b) Digit 6

(c) Absolute area of ACE

Fig. 4: Sample digits (a) 5 and (b) 6. (c) Proposed Metric : Absolute area of ACE "A-ACE" for handpicked pixels



(a) Digit 0 (b) Digit 9

(c) Absolute area of ACE

Fig. 5: Sample digits (a) 0 and (b) 9. (c) Proposed Metric : Absolute area of ACE "A-ACE" for handpicked pixels

References

[1] J. Chen, Z. Xu, and M. Tomizuka, "End-to-end autonomous driving perception with sequential latent representation learning," 2020.

[2] A. Khademi and V. Honavar, *A causal lens for peeking into black box predictive models: Predictive model interpretation via causal attribution*, 2020. arXiv: 2008.00357 [cs.LG].

[3] Y. LeCun, C. Cortes, and C. Burges, "The mnist database.," 2020. [Online]. Available: http://yann.lecun.com/exdb/mnist/.

[4] R Moraffah, M Karami, R Guo, A Raglin, and H Liu, "Causal interpretability for machine learning – problems, methods and evaluation," 2020. arXiv: 2003.03934 [cs.CV].

[5] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavvaf, and E. A. Fox, "Natural language processing advancements by deep learning: A survey," 2020. arXiv: 2003.01200 [cs.CV].

[6] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," 2019. arXiv: 1910.10045 [cs.CV].

[7] A Chattopadhyay, P Manupriya, A Sarkar, and V. Balasubramanian, "Neural network attributions: A causal perspective," *Proceedings of the 36th International Conference on Machine Learning*, 2019. arXiv: 1902.02302 [cs.CV].

[8] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z Feng, and R. Qu, "A survey of deep learning-based object detection," 2019.

[9] Y. Li, J. Hu, Y. Wang, J. Zhou, L. Zhang, and Z. Liu, "Deepscaffold: A comprehensive tool for scaffold-based de novo drug discovery using deep learning," 2019. arXiv: 1908.07209 [cs.CV].

[10] J Pearl, "The seven tools of causal inference, with reflections on machine learning," *Commun. ACM*, vol. 62, no. 3, pp. 54–60, 2019. DOI: 10.1145/3241036. [Online]. Available: https://doi.org/10.1145/3241036.

[11] V. Prabhu, A. Kannan, G. J. Tso, N. Katariya, M. Chablani, D. Sontag, and X. Amatriain, "Open set medical diagnosis," 2019. arXiv: 1910.02830 [cs.CV].

[12] C. Zheng, R. Dai, and M.-J. Zhang, *On high dimensional covariate adjustment for estimating causal effects in randomized trials with survival outcomes*, 2019. arXiv: 1812.02130 [stat.ME].

[13] K. R. Mopuri, U. Garg, and R. V. Babu, *Cnn fixations: An unraveling approach to visualize the discriminative image regions*, 2018. arXiv: 1708.06670 [cs.CV].

[14] T. Nomura and S. Hattori, "Estimation of the average causal effect via multiple propensity score stratification," *Communications in Statistics - Simulation and Computation*, vol. 47, no. 1, pp. 48–62, 2018. DOI: 10.1080/03610918.2016.1208230. eprint: https://doi.org/10.1080/03610918.2016.1208230. [Online]. Available: https://doi.org/10.1080/03610918.2016.1208230.

[15] D Alvarez-Melis and T. S. Jaakkola, "A causal framework for explaining the predictions of black-box sequence-to-sequence models," 2017. arXiv: 1602.04938 [cs.CV].

[16] M. Kocaoglu, C. Snyder, A. G. Dimakis, and S. C. Vishwanath, "Learning causal implicit generative models with adversarial training," 2017. arXiv: 1709.02023 [cs.CV].

[17] K. Raza and S. Ahmad, "Recent advancement in next generation sequencing techniques and its computational analysis," 2017.

[18] A Shrikumar, P Greenside, and A Kundaje, "Learning important features through propagating activation differences," *Proceedings of the 34th International Conference on Machine Learning*, 2017. eprint: 1704.02685.

[19] M Sundararajan, A Taly, and Q Yan, "Axiomatic attribution for deep networks," 2017. arXiv: 1703.01365 [cs.CV].

[20] M. T. Ribeiro, S Singh, and C Guestrin, "Why should i trust you?": Explaining the predictions of any classifier," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. arXiv: 1602.04938 [cs.CV].

[21] S Bach, A Binder, G Montavon, F Klauschen, M. K-R, and Samek, "Pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10(7): e0130140, 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0130140.

[22] O. Russakovsky, J. Deng, H. Su, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision (IJCV)*, vol. 115(3), 2015.

[23] J Pearl, "A probabilistic calculus of actions," Feb. 2013. arXiv: 1302.6835 [cs.CV].

[24] M. Anusuya and S. Katti, "Speech recognition by machine: A review," 2009. arXiv: 1001.2267 [cs.CV].

[25] J. Pearl, *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York. 2nd edition, 2009, 2000a.

[26] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search 2nd ed.* MIT Press, Cambridge, MA, 2000.

[27] J. Pearl, "Causal diagrams for empirical research," *Biometrika*, vol. 82, no. 4, pp. 669–688, Dec. 1995, ISSN: 0006-3444. DOI: 10.1093/biomet/82.4.669. eprint: https://academic.oup.com/biomet/article-pdf/82/4/669/698263/82-4-669.pdf. [Online]. Available: https://doi.org/10.1093/biomet/82.4.669.

[28] H. S., "Shapley value," in *Game Theory. The New Palgrave.* J. Eatwell, M. Milgate, and P. Newman, Eds., Springer: Palgrave Macmillan, London, 1989, pp. 210–216.