1 Proof of concept using simple CNN model - Extension to complex models

To illustrate the idea of causal reasoning using the proposed metric "A-ACE", a simple CNN model has been used, as in [1]. As pilot experiment, images from the popular MNIST data have been used. This idea has been extended to include more complex models such as ResNet-101 for images with greater complexity. The results are shown in figure 5 and 6. It must be noted that in order to accomplish this the following tweaks are required.

1. Shape of input data.
2. Shape of mean and covariance.
3. Shape of gradient and hessian.

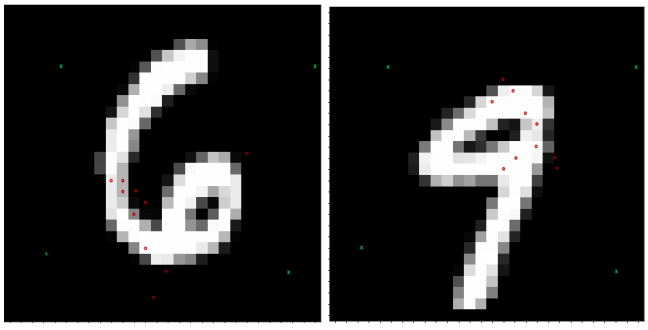
Hence, the experiments are not necessarily limited to simple models alone; complex models can easily be accomodated. Besides, not limiting experiments to simple binary images of digits from MNIST data, we have also shown the performance of the propsed metric on real world images from Imagenet dataset (Review 1E4F)

2 Number of training images - Number of testing images

For the binary classification problem, we have used 30000 training images and 12000 images for testing. Hence, making it 42000 images in all. We have handpicked pixels from the three following regions to interpret the A-ACE measures. 1. Pixels common to both the classes. 2. Background pixels. 3. Class specific pixels. Of these, the A-ACE measure for class specific pixels are shown to be the highest leading to discriminability between the two classes. Of the ten digits, pair-wise classification problems were carried out. Sample results of these are shown on the following pairs.

1. 3 vs 8
2. 6 vs 9
3. 2 vs 5
4. 2 vs 6

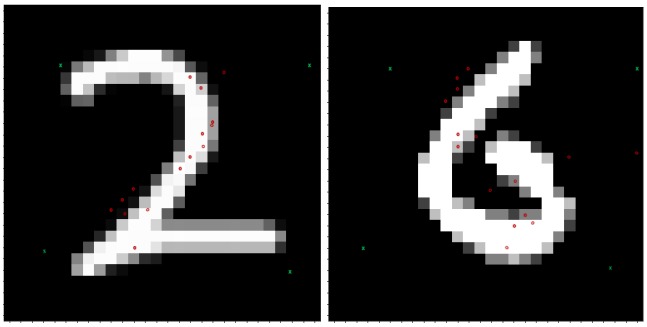
Figure 1-4 illustrate the results on the above digit pairs. Detailed results can be seen here



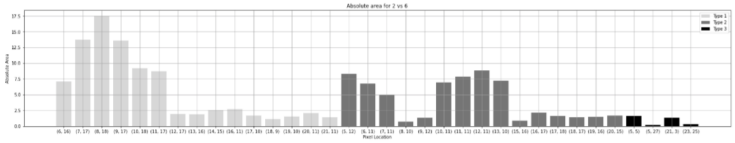
(a) Digit 6 (b) Digit 9

(c) Absolute area of ACE

Figure 1: Sample digits (a) 6 and (b) 9. (c) Proposed Metric : Absolute area of ACE "A-ACE" for handpicked pixels.

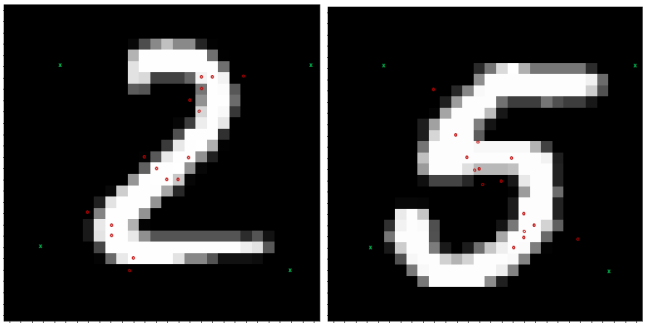


(a) Digit 2 (b) Digit 6

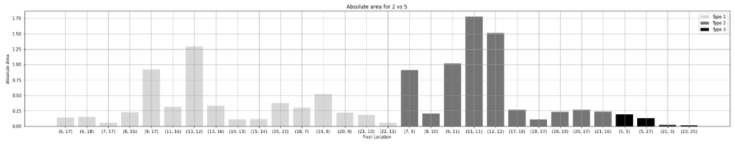


(c) Absolute area of ACE

Figure 2: Sample digits (a) 2 and (b) 6. (c) Proposed Metric : Absolute area of ACE "A-ACE" for handpicked pixels

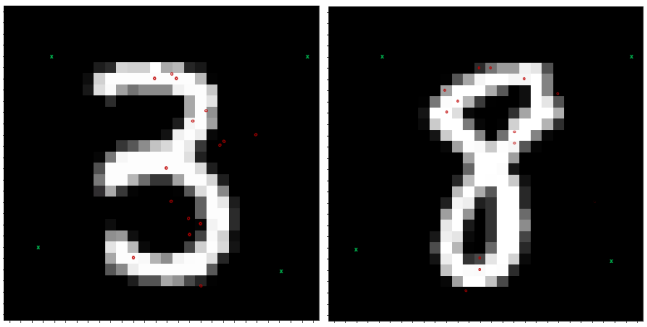


(a) Digit 2 (b) Digit 5

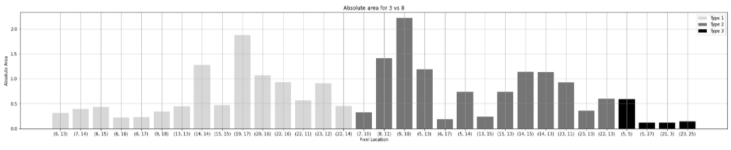


(c) Absolute area of ACE

Figure 3: Sample digits (a) 2 and (b) 5. (c) Proposed Metric : Absolute area of ACE "A-ACE" for handpicked pixels



(a) Digit 3 (b) Digit 8



(c) Absolute area of ACE

Figure 4: Sample digits (a) 3 and (b) 8. (c) Proposed Metric : Absolute area of ACE "A-ACE" for handpicked pixels

Case	% increase in A-ACE peak in distinctive region
3 vs 8	38%
6 vs 9	45%
2 vs 5	33%
2 vs 6	50%

3 Usefulness of A-ACE as compared to ACE

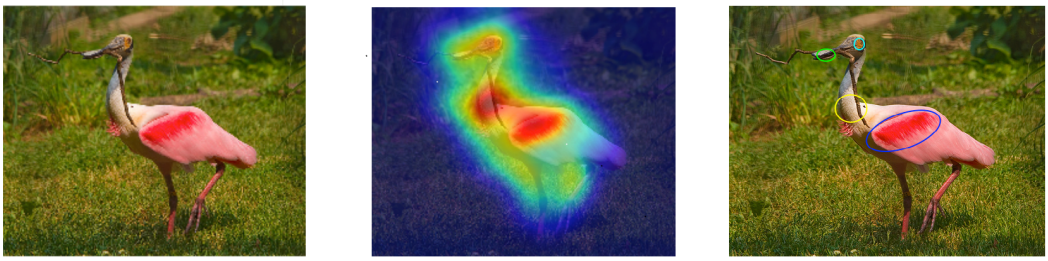
In a binary classification scenario, interpretability is possible only by localizing discriminative regions. In our experiements, we show that the behaviour at the discriminative regions can be captured using the proposed A-ACE. A-ACE exploits the magnitude of the causal effect irrespective of the direction. This leads to improved quantification of interpretability. The limitation in ACE is that it doesn't exploit the variation with changing levels of interventional values.

4 Real world complex images and comparison with SOTA

We have compared proposed method with SOTA (CNN fixation), whose representative result (taken from ILSVRC validation set) is shown in figure 5 and 6, using pretrained ResNet-101 model. Figure 5(a) shows the input image, 5(b) shows the localization map using CNN fixation (SOTA) method and 5(c) illustrates the distinctive regions as output by the proposed method. The distinctive regions turn out to be the feathers, beak, crown, and the neck of the bird which indeed lead to the resulting label.

5 Claim of Interpretability

In a binary classification problem, we have incorporated three type of regions (handpicked) - pixels common to both classes, distinguishing pixels and background pixels for interpreting the model using A-ACE. We consistently find peak at distinguishing pixels to be at least 33% higher than other as shown in figure 1-4.



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

Figure 5: Comparison of the localization regions for sample images from ILSVRC validation set for CNN fixation and proposed metric for ResNet-101. Our metric shows higher A-ACE value for the encircled region in the descending order encoded by color - violet, yellow, cyan and lime

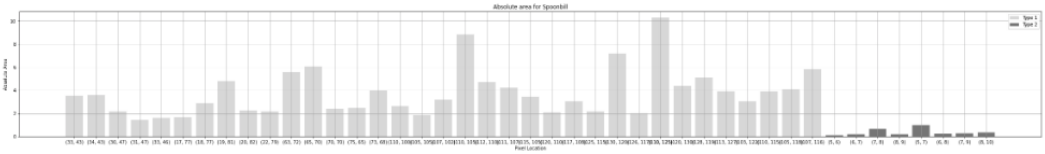


Figure 6: Absolute area of ACE for handpicked pixels of input image

References

[1] 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].