Attribution-Based Confidence Metric For Deep Neural Networks

Susmit Jha
Computer Science Laboratory
SRI International

Sunny Raj, Steven Lawrence Fernandes, Sumit Kumar Jha
Computer Science Department
University of Central Florida, Orlando

Somesh Jha
University of Wisconsin-Madison
and Xaipient

Brian Jalaian, Gunjan Verma, Ananthram Swami
US Army Research Laboratory
Adelphi

Attributions

Additive feature attribution locally: Boolean features - present or absent

$$g(x) = a_0 + \sum_i^M a_i x^i$$

From cooperative game theory, we have classic equations to compute Shapley values

$$a_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}\left(x_{S \cup \{i\}}\right) - f_S(x_S) \right]$$

- **Local accuracy**: $g(x) = f(x)$ [explanation model matches original model on the input]
- **Sensitivity**: $x^i = 0 \Rightarrow a_i = 0$
- **Consistency**: For any two models f and f', $f'(x) - f'\left(x \setminus \{x^i\}\right) \geq f(x) - f\left(x \setminus \{x^i\}\right)$ for all presence/absence of features x in , then $a_i^{f'} \geq a_i^f$

Young (1985) demonstrated that Shapley values are the only set of values that satisfy these properties.

$$a_i = \sum_{z \subseteq x} \frac{|z|!\,(M - |z| - 1)!}{M!} \left[f_x(z) - f_x(z' \setminus \{x^i\}) \right]$$

Apply sampling approximations to above equation and approximate the effect of removing a variable from the model by integrating over samples.

Shapley Values

Young (1985) demonstrated that Shapley values are the only set of values that satisfy these properties.

$$a_i = \sum_{z \subseteq x} \frac{|z|!\,(M - |z| - 1)!}{M!} \left[f_x(z) - f_x(z' \setminus \{x^i\}) \right]$$

Apply sampling approximations to above equation and approximate the effect of removing a variable from the model by integrating over samples.

Baseline and path based methods.

Friedman, Eric J. Paths and consistency in additive cost sharing. International Journal of Game Theory, 32(4): 501–518, 2004.

Given $\gamma = (\gamma_1, \ldots, \gamma_n)$: $[0,1] \rightarrow R^n$ be a smooth function specifying a path in R^n from baseline x^b to input x, that is, $\gamma(0) = x^b, \gamma(1) = x$.

$$\int_{\alpha=0}^{1} \frac{\partial F(\gamma(\alpha))}{\partial \gamma_i(\alpha)} \frac{\partial \gamma_i(\alpha)}{\partial \alpha} \, d\alpha$$

$$\mathcal{A}_j^i(\mathbf{x}) = (\mathbf{x}_j - \mathbf{x}_j^b) \times \int_{\alpha=0}^{1} \partial_j \mathcal{F}^i(\mathbf{x}^b + \alpha(\mathbf{x} - \mathbf{x}^b)) d\alpha$$

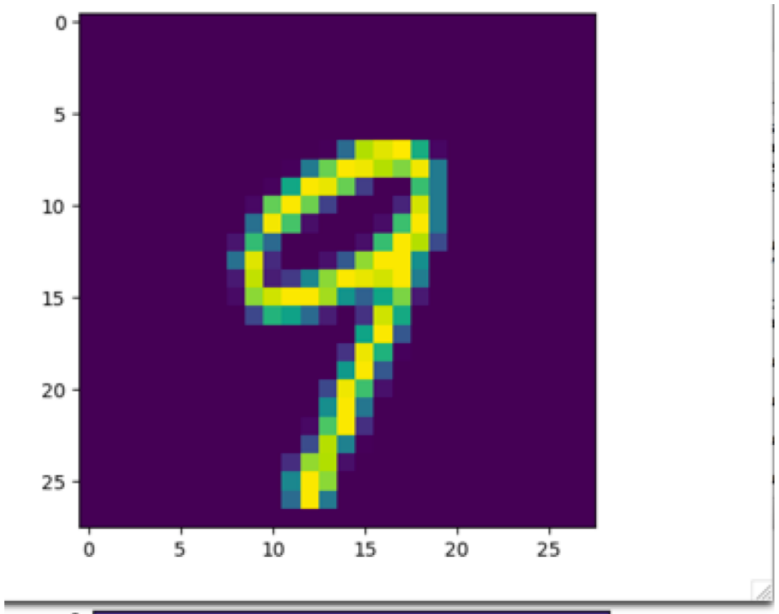
IG. Sundararajan et. al.'17

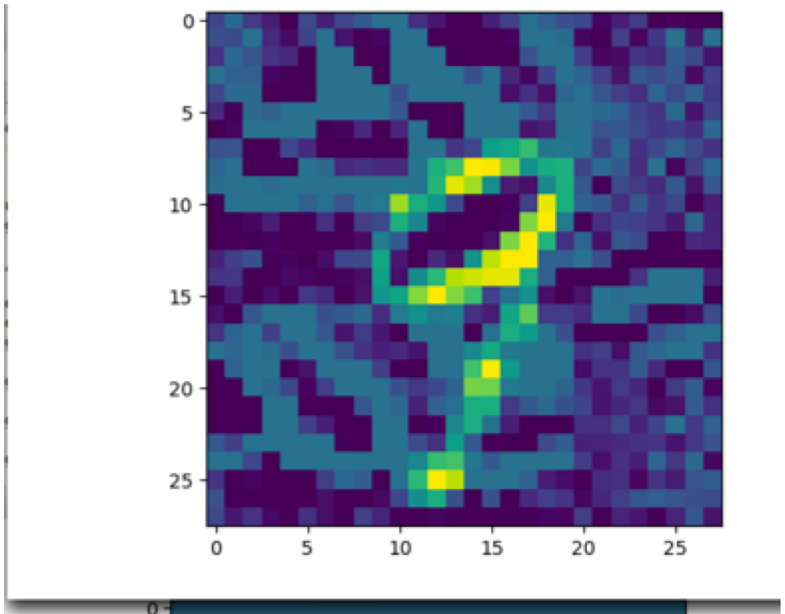
Compositional

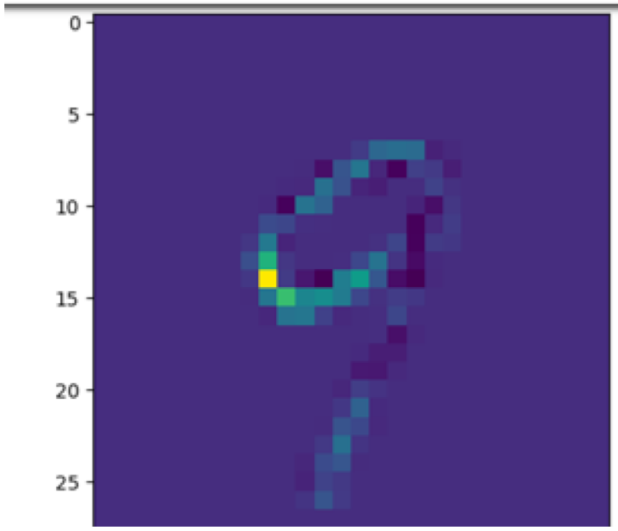
$$f(x) = \sum_{j=1}^{M} w_j x_j + b$$

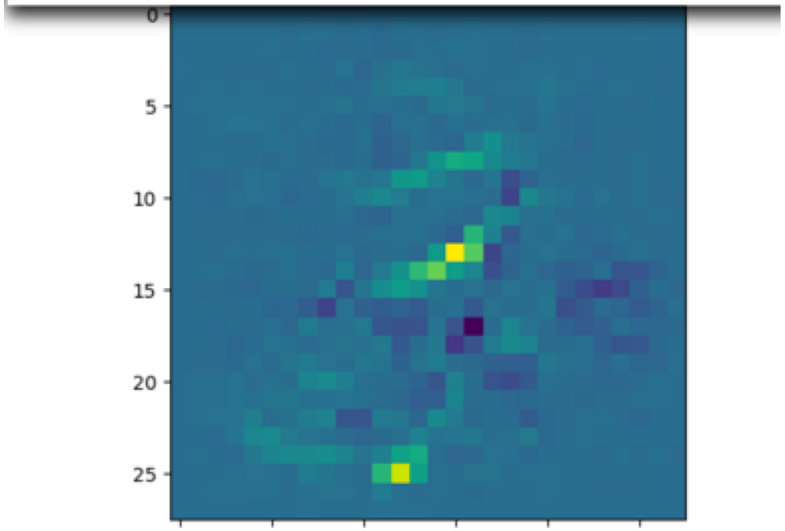
$$a_i = \mathrm{w_i}\,(\mathrm{x_i} - \mathrm{E}[\mathrm{x_i}])$$

Sample Attributions (9 classified as 9, noisy 9 classified as 3)

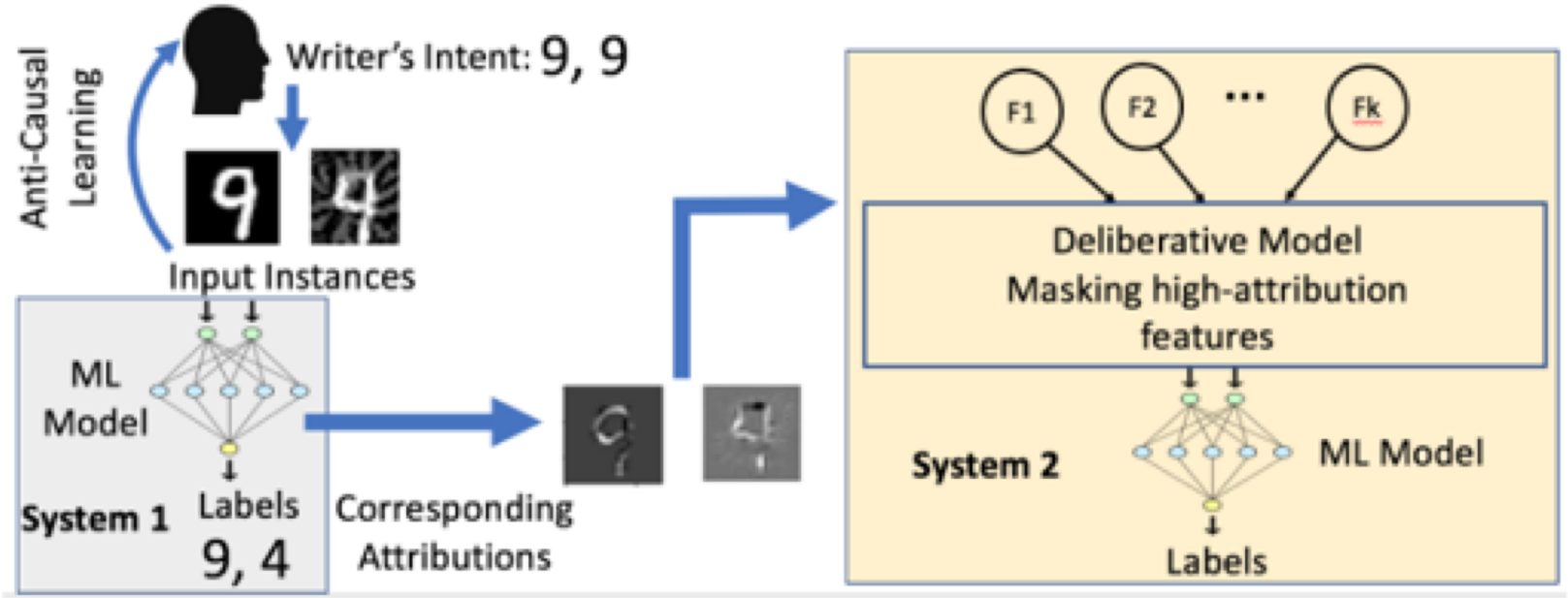








ABC: Attribution Based Confidence



1. If the model makes a prediction on an input robustly in the causal neighborhood, that is, there is redundancy of features on an input, then it is more confident.
2. For out-of-distribution or adversarial examples, the model's prediction is not robust in causal space.

Attribution Based Confidence (ABC)

Given an input $\mathbf{x}$ for a model $\mathcal{F}$ where $\mathcal{F}_i$ denotes the i-th logit output of the model, we can compute attribution of feature $\mathbf{x}_j$ of $\mathbf{x}$ for label i as $\mathcal{A}_j^i(\mathbf{x})$. We can then obtain confidence in two steps:

- Sample in neighborhood of $\mathbf{x}$ by mutating each feature $\mathbf{x}_j$ with probability $\frac{|\mathcal{A}_j^i(\mathbf{x})/\mathbf{x}_j|}{\sum_j |\mathcal{A}_j^i(\mathbf{x})/\mathbf{x}_j|}$ where the feature $\mathbf{x}_j$ is changed to flip the label away from i.
- Report the fraction of samples points in the neighborhood of input $\mathbf{x}$ for which the decision of the model conforms to the original decision as the conservatively estimated confidence measure.

Algorithm 1 Evaluate confidence $c(\mathcal{F}, \mathbf{x})$ of machine learning model $\mathcal{F}$ on input $\mathbf{x}$

Input: Model $\mathcal{F}$, Input $\mathbf{x}$ with features $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n$, Sample size S
Output: Confidence metric $c(\mathcal{F}, \mathbf{x})$

1: $\mathcal{A}_1, \ldots \mathcal{A}_n \leftarrow$ IG attributions of features $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n$ from input $\mathbf{x}$
2: $i \leftarrow \mathcal{F}(\mathbf{x})$ {Obtain model prediction}
3: **for** $j = 1$ to n **do**
4: $\quad P(\mathbf{x}_j) \leftarrow \frac{|\mathcal{A}_j/\mathbf{x}_j|}{\sum_{k=1}^{n} |\mathcal{A}_k/\mathbf{x}_k|}$
5: **end for**
6: Generate S samples by mutating feature $\mathbf{x}_j$ of input $\mathbf{x}$ to baseline with probability $P(\mathbf{x}_j)$
7: Obtain the output of the model on the S samples.
8: $c(\mathcal{F}, \mathbf{x}) \leftarrow S_{conform}/S$ where model's output on $S_{conform}$ samples is i
9: **return** $c(\mathcal{F}, \mathbf{x})$ as confidence of prediction by the model $\mathcal{F}$ on the input $\mathbf{x}$

Attribution with Adversarial Attacks

Original image with a label of yawl

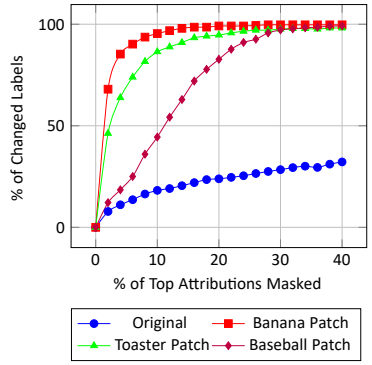
Masking its top 0.2% of attribution

Masking its top 0.4% of attribution

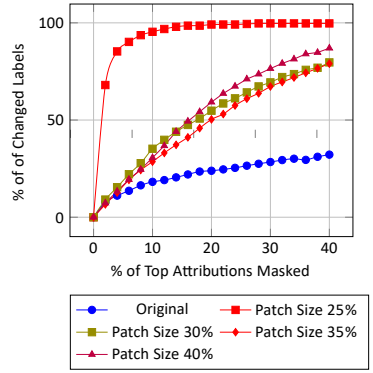
Image with a banana patch generated using adversarial patch method

Masking its top 0.2% of attribution

Masking its top 0.4% of attribution

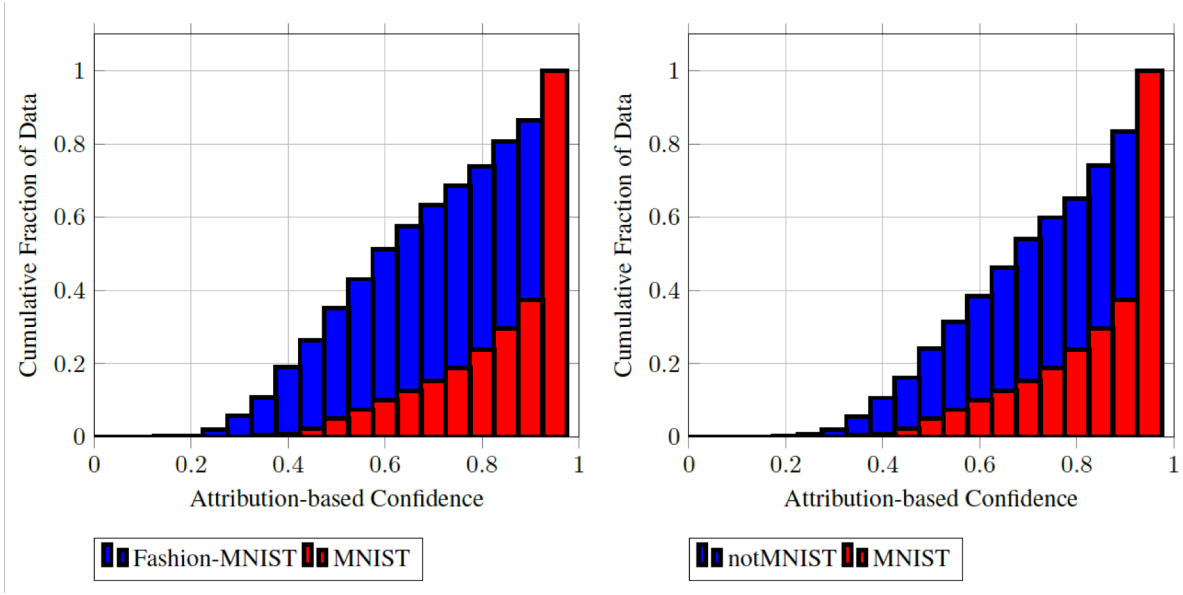


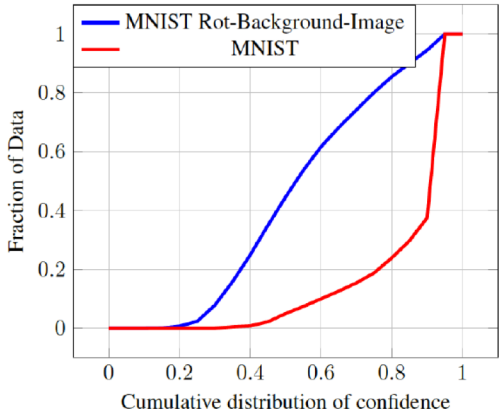
Dropping 0.4% of the attribution causes 99.71% of the attacks based on banana patches, 98.14% of the attacks based on toaster patches, and 99.20% of the attacks based on baseball patches to be detected.

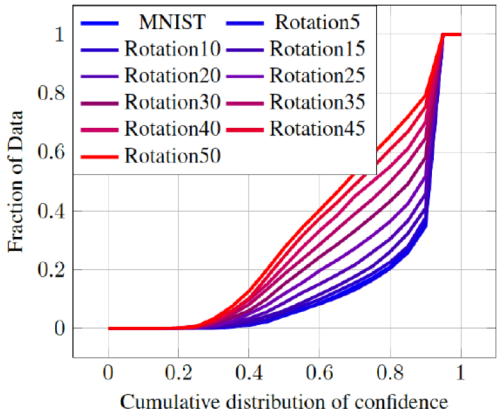


Masking 0.4% of attributions caused nearly 80% of labels to change for images with adversarial patches.

ABC: Attribution Based Confidence (MNIST)





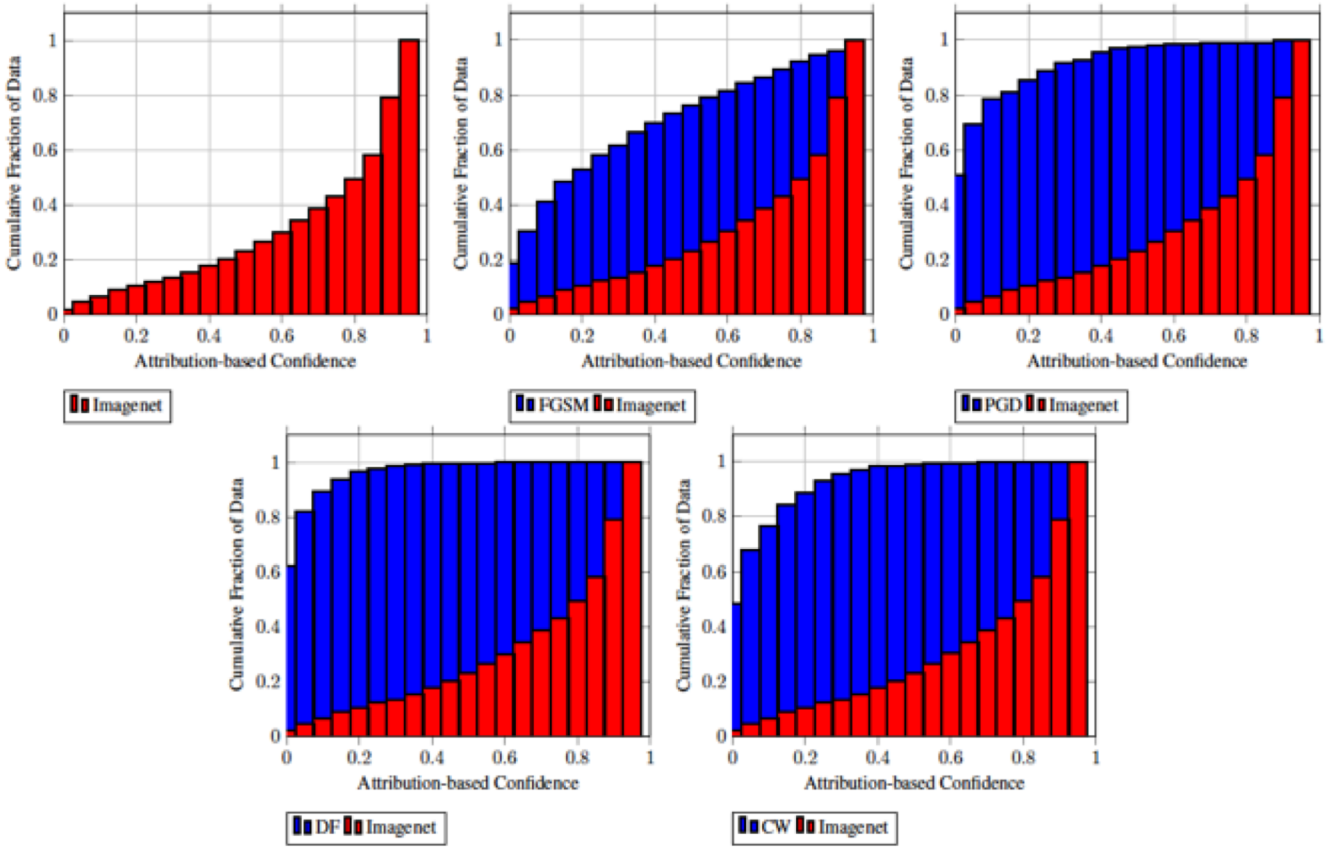


2 misclassified as 9
AttributeConf=0.28

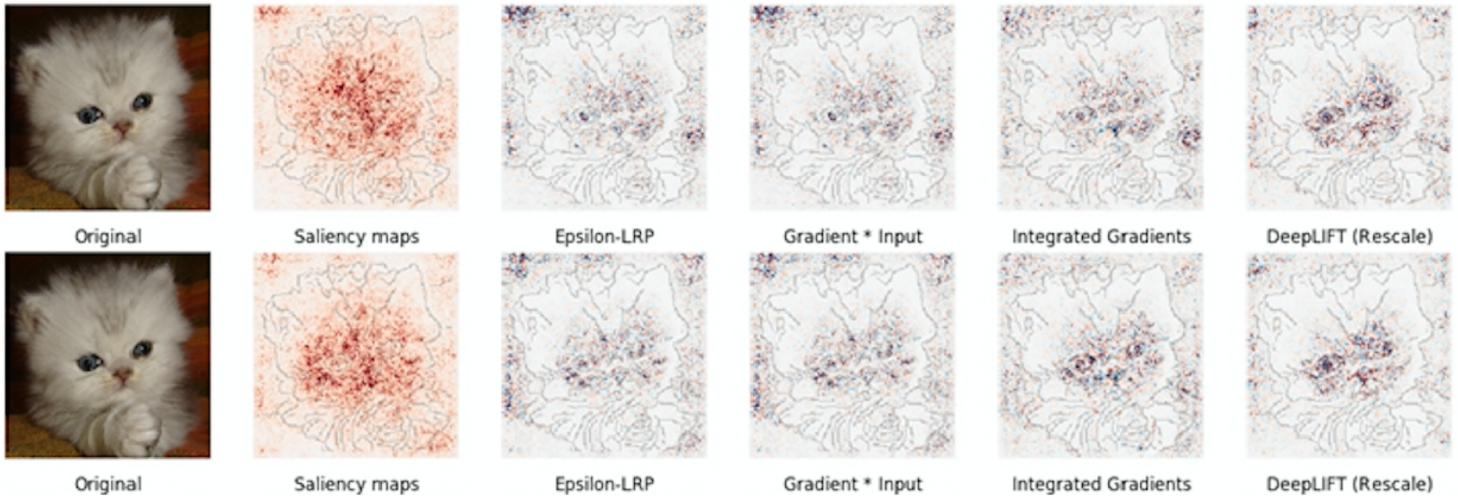
3 misclassified as 2
AttributeConf=0.41

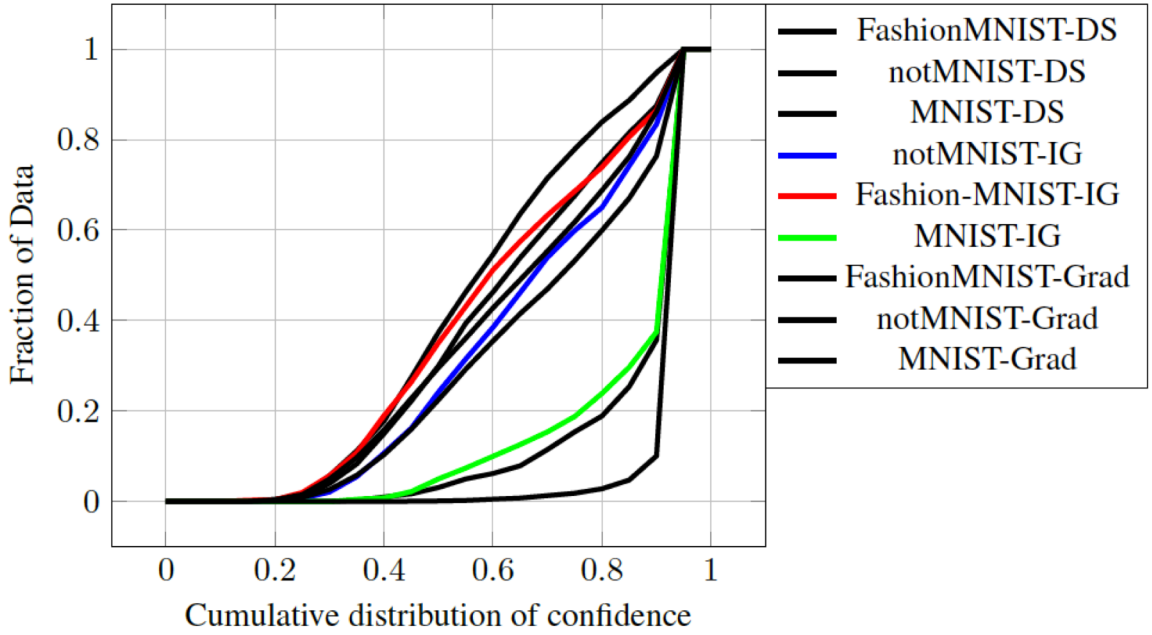
8 misclassified as 2
AttributeConf=0.89

ABC: Attribution Based Confidence (ImageNet)



ABC: Different Attribution Methods





Conclusion

We propose a novel attribution-based confidence (ABC) metric computed by importance sampling in the neighborhood of a high-dimensional input using relative feature attributions, and estimating conformance of the model. It does not require access to training data or additional calibration.

We empirically evaluate the ABC metric over MNIST and ImageNet datasets using

(a) out-of-distribution data,
(b) adversarial inputs generated using digital attacks such as FGSM, PGD, CW and DeepFool, and
(c) physically-realizable adversarial patches and LaVAN attacks.