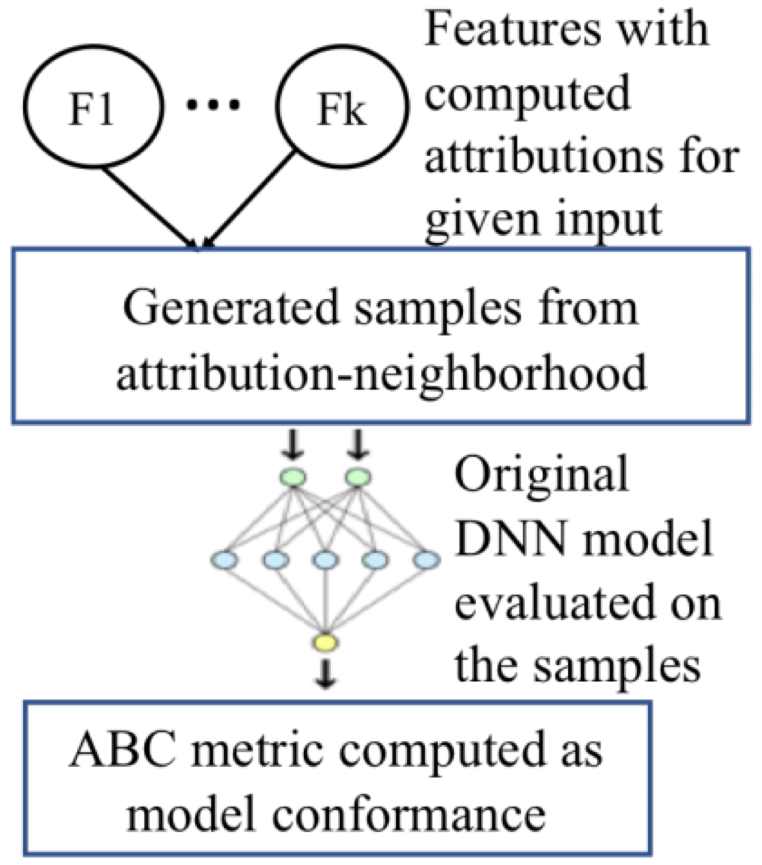
CPS:Small: Self-Improving Cyber-Physical System.
Attribution-Based Confidence (ABC) for Deep Learning Models

Susmit Jha, SRI
http://csl.sri.com/users/jha/projects/si-cps/sicyps.html

The project pursues the goal of developing the science for designing safe, yet optimal, active data-driven adaptive cyber-physical systems. This requires development of data-driven learning techniques that can quantify uncertainty in prediction and report this confidence measure. The rest of CPS will use the learning model's output and its confidence via uncertainty-aware control.

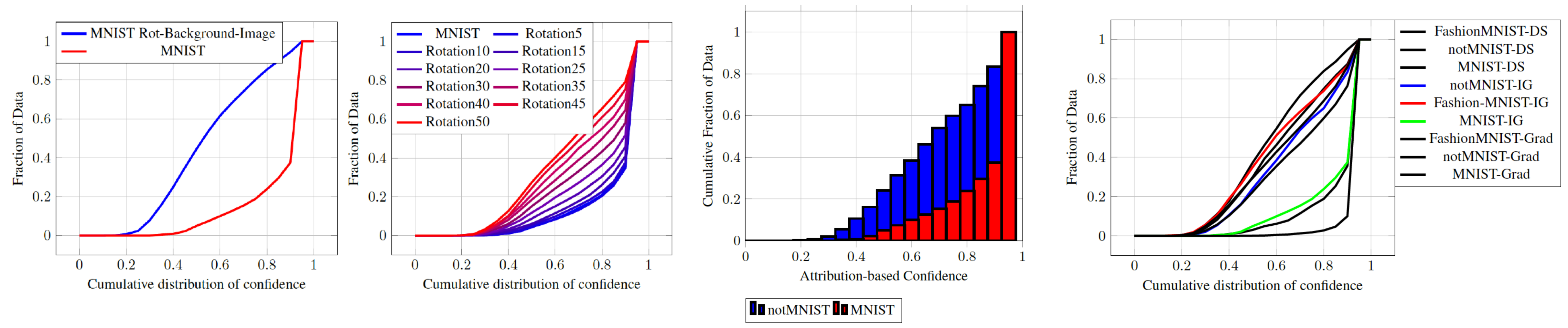
Focus of the poster: Attribution-based confidence of learned models.



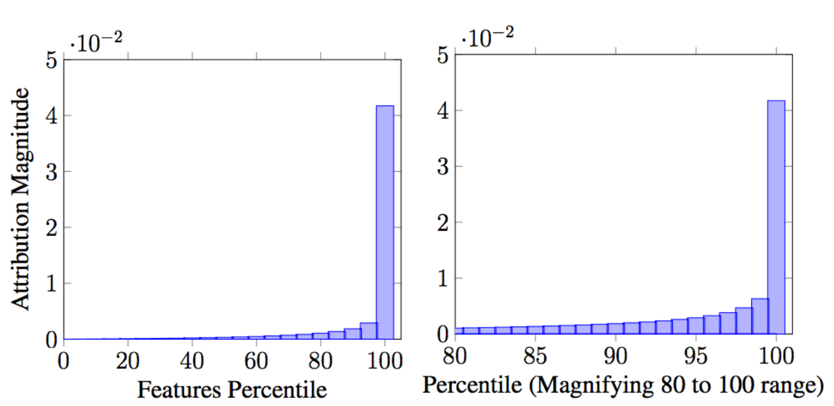
Given an input $\mathbf{x}$ for a model $\mathcal{F}$ where $\mathcal{F}_i$ denotes the i-th logit output of the model, we can compute attribution of feature $\mathbf{x}_j$ of $\mathbf{x}$ for label i as $\mathcal{A}^i_j(\mathbf{x})$. We can then obtain confidence in two steps:

- Sample in neighborhood of $\mathbf{x}$ by mutating each feature $\mathbf{x}_j$ with probability $\frac{|\mathcal{A}^i_j(\mathbf{x})/\mathbf{x}_j|}{\sum_j |\mathcal{A}^i_j(\mathbf{x})/\mathbf{x}_j|}$ 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.

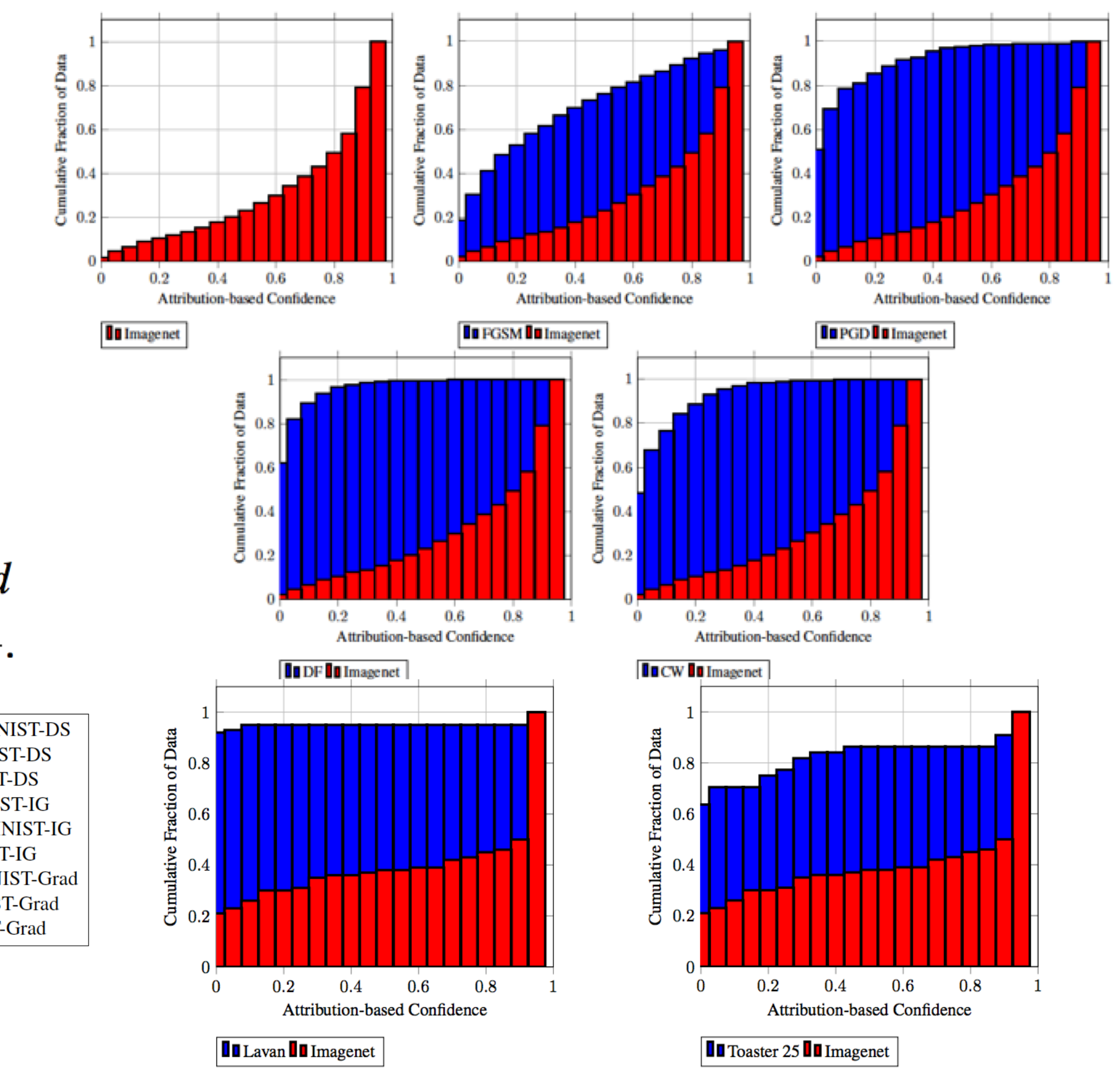
Theorem 1. *The sensitivity of the output $\mathcal{F}(\mathbf{x})$ with respect to an input feature $\mathbf{x}_j$ in the neighborhood of $\mathbf{x}$ is approximately the ratio of the attribution $\mathcal{A}_j(\mathbf{x})$ to the value of that feature $\mathbf{x}_j$, that is, $\frac{\mathcal{A}_j(\mathbf{x})}{\mathbf{x}_j}$.*

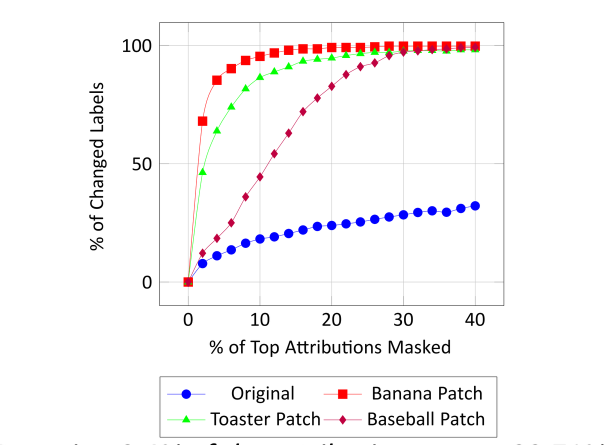


Out of distribution data leads to low confidence prediction

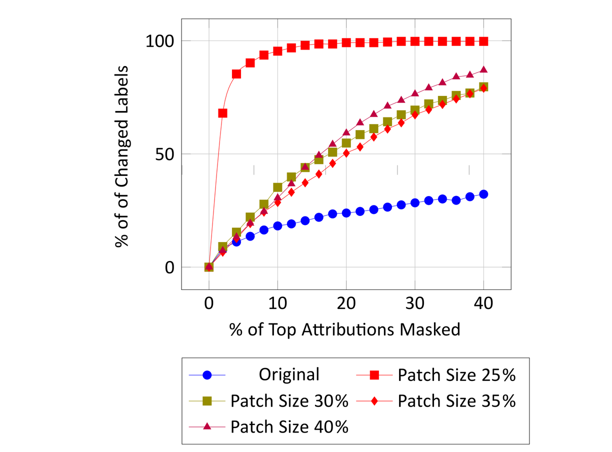


Feature Concentration in well-trained models





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.

Adversarial attacks lead to low confidence prediction

Summary

We proposed a novel attribution-based confidence (ABC) metric.. It does not require access to training data or additional calibration.
We empirically evaluated 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.

Reference: Jha et. al. Attribution-Based Confidence (ABC) Metric For Deep Neural Networks. Thirty-third Conference on Neural Information Processing Systems (NeurIPS), 2019

Other research results on the project this year:

Logic extraction thrust led publications in JAR'18, NeurIPS'18, FMSD'19, AAAI Consciousness Symposium'19 and a tutorial at NSV'19 (co-located with CAV'19)

Uncertainty and risk-aware control thrust led to research results published in JAR'18, Allerton'18, American Control Conference'19, and HSCC'19.

Broader Impact:

3 student interns were supported in part by this project this year. 1 of the 3 students was a woman student.

SRI started a collaboration with CodeChix - a non-profit focused on the retention of women in technology (https://www.codechix.org/). Presented at first joint summit in September, 2019.

2019 NSF Cyber-Physical Systems Principal Investigators' Meeting
November 21-22, 2019 | Crystal City, Virginia

Award ID#: CNS CPS:Small 1740079