Causal Neighborhood Analysis for Detection of Adversarial Examples

Sunny Raj, Sumit Kumar Jha
Computer Science Department
University of Central Florida, Orlando
{sraj,jha}@eecs.ucf.edu

Susmit Jha
Computer Science Laboratory
SRI International
susmit.jha@sri.com

Abstract

Attribution methods have been developed to explain the decision of a machine learning model on a given input. These methods identify quantitative attribution of different features of an input for a decision made by a machine learning (ML) model. This enables top-down causal inference from features to input after a bottom-up classification decision has been made by an ML model. We define the causal neighborhood of an input by incrementally masking high attribution features, and generating new inputs. While bottom-up classification models have been shown to be brittle and non-robust, the top-down causal inference presents an effective approach to detect adversarial attacks or unexpected perturbations. In this paper, we study the robustness of machine learning models on benign and adversarial inputs in its causal neighborhood. We experimentally show that ML models are robust in the causal neighborhood of the benign inputs but the adversarial inputs rely on few high attribution features, and consequently lack robustness in their causal neighborhood. We demonstrate our results on the state-of-the-art adversarial attack methods such as DeepFool, FGSM, CW and PGD, as well as physically realizable adversarial examples. Our results indicate that heavy concentration of attribution over few features in physically realizable adversarial examples, makes our approach specially suitable for detecting them. Such a defense approach is independent of training data and attack method, and is a first step towards the use of causal inference in making ML models resilient.

1 Introduction

The adoption of deep learning systems in safety-critical or high-security applications is inhibited due to two major concerns: their brittleness to adversarial attack methods that can make imperceptible modification to inputs and trigger wrong decisions [1, 2], and the lack of interpretability [3]. Significant progress has also been made towards adversarial robustness [4–6] and explainability [7–9]. In this paper, we investigate this connection between the resilience to adversarial perturbations and the attribution-based explanation of individual decisions by a machine learning model. Our study is motivated by recent arguments on robustness of causal learning [10], the use of causal inference to compute the confidence of a machine-learned model on a given input [11], and Kahneman's decomposition [12] of cognition into two layers: System 1, or the intuitive system learning an anti-causal model (from input to the label/intent), and System 2, or the deliberate system doing causal inference (from intent to input). The original machine learning model represents the System 1 and the attribution analysis presented in this paper is the deliberative System 2 that can detect inconsistencies in System 1 by analyzing the attribution generated by the model on the input. Our central hypothesis is that adversarial inputs are imperceptibly similar to benign inputs and are still able to trigger wrong decisions because of a relatively small number of features with high attribution

Workshop on Perception as Generative Reasoning, NeurIPS 2019, Vancouver, Canada.

Figure 1: (From left to right) original image; masking its top 0.2% attribution; masking its top 2% of attribution; image with a baseball patch generated using LaVAN method; masking its top 0.2% attribution; (rightmost) masking top 2% of its top attributions. Most pixels are masked from the noise patch leading to a change in label. Original images are robust to pixel masking and retain their labels.

in the machine learning model. This concentration of causal attributions is central to their physical realizability and perceptual indistinguishably from benign inputs.

Adversarial perturbations often create a small fraction of high-attribution features responsible for the change in the output model without visible change in the input. Consequently, it is possible to identify an adversarial example by examining inputs in its causal neighborhood obtained by incrementally masking the features which have high magnitude attributions. While benign inputs are robust and the model does not change its decision on the input in this neighborhood, the decision of the model is not robust on the adversarial examples. The localized nature of existing physically realizable attacks naturally concentrates the high attribution features. The ease of detection of these adversarial examples suggests a trade off in physical realizability and indistinguishability in attribution space. In this paper, we propose a defense based on Kahneman's decomposition of cognition into intuitive System 1 and deliberative System 2. This approach does not rely on analyzing training data such as manifold-based defense [13, 14], or statistical signature of the machine learning models such as logit pairing [6], or methods that exploit the knowledge of specific attack for adversarial training [15].

2 Background

Adversarial examples [16] which produce incorrect decision from machine learning models can be obtained using a number of techniques [17, 5, 18–20] which are now available in tools such as Cleverhans [2]. While these methods rely on diffused digital transformation of the input, physically realizable attacks in form of patches or stickers that can be added to an image have also been developed. Adversarial patch [21] and localized and visible adversarial noise(LaVAN) [22] methods generate patches that are universal and can be used to attack any image and are targeted to force classifier to output a particular output label. Their goal is to generate universal noise "patches" that can be physically printed and put on any image, in either a black-box (when the attacked network is unknown) or white-box (when the attacked network is known) setup. A related attack is proposed in [23] to investigate the blind-spots of state-of-the-art image classifiers, and the kinds of noise that can cause them to misclassify. More diffused universal adversarial digital perturbations [24] have also been proposed that can be applied to any input.

A number of attribution techniques have been recently proposed in literature that assign positive and negative importance to an input feature for a given decision of the machine learning model. Attributions can correspond to the actual decision or to counterfactuals. Many prominent attribution methods are based on the gradient of the predictor function with respect to the input [25, 26, 9]. These approaches have been extended to counterfactual analysis [27] which has roots in cooperative game theory and revenue division. Counterfactual analysis is difficult when the number of possible output labels is large. We restrict our study to analysis of the actual attribution. Different attribution methods are compared in [28]. The sensitivity of these attributions to perturbations in the input are studied in [29] and adversarial attacks are presented to create perceptively indistinguishable images with the same prediction label but different attributions. Their results indicate that both System 1 and System 2 in our proposed approach can be independently attacked.

The goal of this paper is to combine both System 1 and System 2 to create a more resilient cognition model than any one of them alone. We use the attributions to identify the cause of the current decision of the machine learning model, and then construct a generator that can mask high-attribution features to explore the causal neighborhood of the input and observe model's behavior in this neighborhood. While learning is still in the anti-causal direction, we add a layer of causal deliberative System 2 that reasons in forward direction to evaluate the robustness of the obtained attribution.

3 Approach

For a given neural network machine learning model M and input x, the attribution for the model output $M(x)$ is computed for the features $F_1, F_2, \ldots, F_k$ of the input. We adopt a simple masking model for causal generation of new inputs from the features. Given an input x with baseline x', the masking generator G can select a subset of $n \leq k$ features $M = \{m_1, m_2, \ldots m_n\}$ to be masked and generate a new input $G(x, M)$ such that $F_i(G(x, M)) = F_i(x)$ for $i \notin M$ and $F_i(G(x, M)) = F_i(x')$ for $i \in M$. For example, if the baseline is a dark image and the features are pixels of the image, the masking model selects a set of pixels M of an input image and makes them dark. We use the attribution over features and this masking model to define a causal neighborhood of the input x as:

Definition: Given an input x with features $F_1, F_2, \ldots, F_k$ and the corresponding attributions $IG(F_1(x))$, $IG(F_2(x))$, $\ldots, IG(F_k(x))$, we sort the features in decreasing order of their attribution $F_{j_1}, F_{j_2}, \ldots, F_{j_k}$ such that the attributions $IG(F_{j_1}(x)) \geq IG(F_{j_2}(x)) \geq \ldots \geq IG(F_{j_k}(x))$. We define a δ causal neighborhood of the input x for $0 \leq \delta \leq 1$ by constructing new inputs $G(x, M)$ obtained via masking features $M \subseteq \{F_{j_1}, F_{j_2}, \ldots, F_{j_{\lceil \delta k \rceil}}\}$, that is,

$$\mathcal{N}_\delta(x) = \{G(x, M) | M \subseteq \{F_{j_1}, F_{j_2}, \ldots, F_{j_{\lceil \delta k \rceil}}\}\}$$

An input (benign or adversarial) is δ robust with respect to its attributions if the machine learning model produces the same output on all inputs in $\mathcal{N}_\delta(x)$. We restrict our study to positive attributions. We denote this restricted neighborhood as $\mathcal{N}^+_\delta(x)$ where all features F_{j_i} in M are guaranteed to have positive attributions, that is, $IG(F_{j_i}(x^M)) > 0$ for all $x^M \in \mathcal{N}^+_\delta(x)$. The monotonicity of M with respect to the inclusion of features positive attributes leads to the following theorem:

Theorem: If the outputs of a machine learning model M on an input x and for some input in its causal neighborhood $x^M \in N^+_\delta(x)$ are different, then the model M also produces different outputs on the input x and $G(x, M_\delta)$ where $M_\delta = \{F_{j_1}, F_{j_2}, \ldots, F_{j_{\lceil \delta k \rceil}}\}$.

This monotonicity enables us to check the robustness of a machine learning model on an input in its causal neighborhood by just considering the farthest input. If the model produces different output, then it is not robust on this input. We can lift the computation of the robustness of machine learning model M on input x to define the attribution sensitivity on a dataset X as follows: $\mathcal{S}(M, X, \delta) = 1 - \frac{|\{x | x \in X \text{ and } \rho(M,x) \leq \delta\}|}{|X|}$ where $|\cdot|$ denotes the size of the set. The central goal of the paper is to study the connection between attributions and adversarial robustness, and to argue the need for Kahneman's System 1 and System 2 cognition for more robust perception. We accomplish this by computing the attribution sensitivity on original datasets and adversarially perturbed datasets. An effective adversarial attack must have smaller attribution sensitivity similar to original inputs to avoid detection.

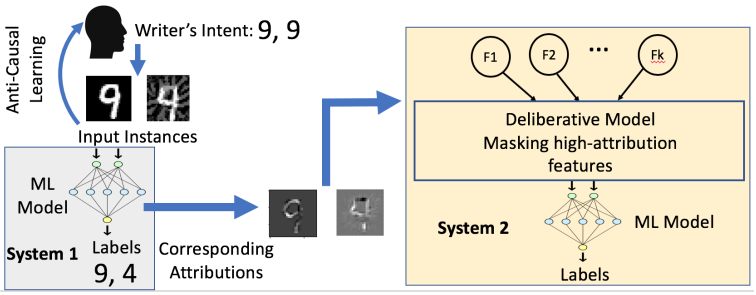

Figure 2: The architecture of the proposed approach motivated by the two level Kahneman's decomposition of cognition. We view the ML model as System 1 and use attribution methods (Integrated Gradient in our experiments) to obtain features with positive and negative attributions. In this example with the MNIST dataset, we see that the adversarial perturbation that causes misclassification of 9 into 4 also significantly changes the attributions. For example, the top part of the perturbed 9 (misclassified as 4) has negative attribution. In deliberative System 2, we perform reasoning in the causal direction, and mask the high attribution features (pixels in this case) to obtain a number of input instances in the causal neighborhood of the original image. The original attributions are robust but the adversarial attributions are not robust which causes the model to assign different labels to images in the causal neighborhood of adversarial examples.

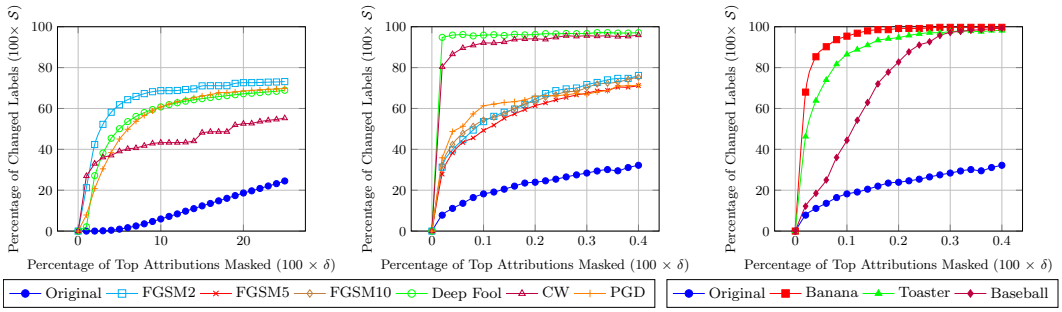

Figure 3: (Left) Percentage of changed labels and percentage of masked attributions for adversarial attacks FGSM, DeepFool, CW, and PGD on MNIST dataset. (Center) Percentage of changed labels and fraction of masked pixels on ImageNet database. FGSM attacks with different $\epsilon = 2, 5, 10$ are evaluated for ImageNet. (Right) Percentage of correct labels and masked attributions for images from the ImageNet dataset is shown in blue. Banana patch and toaster patch were generated using adversarial patch method [21]. Baseball patch was generated using LaVAN method [22].

4 Results

We use Integrated Gradients [9] to obtain attributions to perform attribution sensitivity analysis on three datasets to test our hypothesis that adversarial examples are less robust (more sensitive) to masking in the attribution space when compared to original images:

MNIST Dataset and Digital Attacks: Removing 20% of the attributions does not cause change in the labels assigned to 81% of the MNIST images. In contrast, masking the top 20% attribution has a much stronger impact on the adversarial images. As shown in Figure 3, this changes the assigned labels of adversarial images generated using FGSM, DeepFool, CW and PGD by 72%, 67%, 53%, and 68% respectively.

ImageNet Dataset and Digital Attacks: The masking of a small number of high-attribution pixels does not cause a change in the labels of the original ImageNet images as shown in Figure 3. Labels assigned to 82% of ImageNet images remain unchanged when pixels corresponding to top 0.1% of attributions are masked. The fraction of unchanged labels remains at 67% even when pixels corresponding to top 0.4% of attributions are masked. In contrast, $76\%, 71\%$ and 75% of the images changed label on masking 0.4% of pixels corresponding to top attributions for the adversarial examples generated using FGSM with different epsilon values. This percentage for PGD($\epsilon = 2$) was 71%, 97% for CW and 95% for DeepFool.

ImageNet Dataset and Physical Attacks: Using 1000 images from the ImageNet dataset, we created adversarial patch and LaVAN attacks for the InceptionV3 deep learning model. For adversarial patch attack we used a patch size of 25% for two patch types: banana and toaster. We observe that there was a change in label for 99.71% of images having the banana patch on masking top 0.4% of high attribution pixels. Similarly, 98.14% of images having the toaster patch changed label on masking top 0.4% of high attribution pixels. For LaVAN attack we used baseball patch of size 50×50 pixels. We observe that 99.20% of images having baseball patch changed label when we masked top 0.4% of attributions. Detection percentages for various pixel masking percentages for images with banana, toaster and baseball patch are shown in Figure 3. We observe that most of the high attribution pixels in adversarial images is localized in the patch and masking these pixels leads to a change in the image label. An example of this behaviour can been seen for the LaVAN method in Figure 1, where most of the masking is localized around the baseball patch.

5 Conclusion

We devise a new defense approach for machine learning models that uses attribution and incremental masking for filtering out adversarial examples. We propose such a deliberative masking based analysis approach as System 2 of Kahneman's two layer cognition system with the actual machine learning model serving as System 1. We demonstrate the effectiveness of our approach on a set of benchmarks that include diffused digital perturbations as well as physically realizable perturbations.

References

[1] Christian Szegedy et al. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[2] Nicolas Papernot, I Goodfellow, et al. Cleverhans v2. 0.0: an adversarial machine learning library. *arXiv preprint arXiv:1610.00768*, 2016.

[3] David Gunning. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2017.

[4] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *ISSP'16*, 2016.

[5] Aleksander Madry et al. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[6] Logan Engstrom, Andrew Ilyas, and Anish Athalye. Evaluating and understanding the robustness of adversarial logit pairing. *arXiv preprint arXiv:1807.10272*, 2018.

[7] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *CVPR*, pages 5455–5463, 2015.

[8] Kwang Moo Yi et al. Lift: Learned invariant feature transform. In *ECCV*, pages 467–483. Springer, 2016.

[9] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, pages 3319–3328. JMLR. org, 2017.

[10] Niki Kilbertus, Giambattista Parascandolo, and Bernhard Schölkopf. Generalization in anticausal learning. *arXiv preprint arXiv:1812.00524*, 2018.

[11] Susmit Jha, Sunny Raj, Steven Fernandes, Sumit Kumar Jha, Somesh Jha, Brian Jalaian, Gunjan Verma, and Ananthram Swami. Attribution-based confidence metric for deep neural networks. In *NeurIPS*, 2019.

[12] Daniel Kahneman and Patrick Egan. *Thinking, fast and slow*, volume 1. 2011.

[13] Andrew Ilyas et al. The robust manifold defense: Adversarial training using generative models. *arXiv preprint arXiv:1712.09196*, 2017.

[14] Susmit Jha, Uyeong Jang, Somesh Jha, and Brian Jalaian. Detecting adversarial examples using data manifolds. In *MILCOM*, pages 547–552. IEEE, 2018.

[15] Florian Tramèr et al. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

[16] Ian J et al. Goodfellow. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[17] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*, pages 2574–2582, 2016.

[18] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *ISSP*, pages 39–57. IEEE, 2017.

[19] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9185–9193, 2018.

[20] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

[21] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

[22] Danny Karmon, Daniel Zoran, and Yoav Goldberg. LaVAN: Localized and visible adversarial noise. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2507–2515, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL `http://proceedings.mlr.press/v80/karmon18a.html`.

[23] Danny Karmon, Daniel Zoran, and Yoav Goldberg. Lavan: Localized and visible adversarial noise. *arXiv preprint arXiv:1801.02608*, 2018.

[24] Seyed-Mohsen Moosavi-Dezfooli et al. Universal adversarial perturbations. In *CVPR*, pages 1765–1773, 2017.

[25] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[26] Ramprasaath R Selvaraju et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

[27] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *ISSP*, pages 598–617. IEEE, 2016.

[28] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *NIPS*, pages 9525–9536, 2018.

[29] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. *arXiv preprint arXiv:1710.10547*, 2017.